Kapitel 5
Funktionen mehrerer Veränderlicher

Bislang hatten wir nur Funktionen betrachtet, die von einer einzigen Variablen abhängen. Für die meisten Anwendungen ist das zu viel zu speziell: Egal ob wir ein wirtschaftliches, soziales oder technisches System beschreiben wollen, können wir sicher sein, daß es von einer Vielzahl verschiedener Größen abhängt.

Wir könnten versuchen, dieses Problem zu umgehen, indem wir alle diese Größen mit einer Ausnahme konstant halten und die so entstehende Funktion einer Veränderlichen mit den uns bekannten Methoden untersuchen – dieser *ceteris paribus* Ansatz *(das Übrige ist gleich)* wird in der Tat bei manchen volkswirtschaftlichen Problemen gerne verwendet. Er ist aber sicherlich nicht allgemein einsetzbar, denn die eine variable Größe kann je nach Werten der festgehaltenen Variablen sehr unterschiedliche Effekte haben: Mehrproduktion kann beispielsweise ja nach Zustand des Marktes mal zu Gewinnen, mal zu Ladenhütern führen. Hinzu kommt, daß die Kenngrößen eines Systems selten unabhängig voneinander sind und daher Veränderungen einer Größe oft zwangsläufig zu Veränderungen weiterer Größen führen.

Trotz dieser Schwierigkeiten wollen wir die bewährten Methoden aus der Analysis einer Veränderlichen soweit wie möglich weiter benutzen; wie sich zeigen wird, ist das auch möglich, allerdings müssen sie durch zusätzliche Hilfsmittel ergänzt werden.

Ein wesentliches solches Werkzeug sind, wie schon bei Funktionen einer Veränderlichen, *lineare* Funktionen; teilweise werden wir hier im Mehrdimensionalen daher auch Methoden aus der Linearen Algebra brauchen.

§1: Grundlegende Eigenschaften

Bevor wir zur Differential- und Integralrechnung im $\mathbb{R}^n$ kommen, brauchen wir – wie schon im eindimensionalen Fall – einige Vorbereitungen über Konvergenz, Stetigkeit und ähnliche Grundbegriffe. Diese sollen hier zusammengestellt werden.

a) Visualisierung in höheren Dimensionen

Um einen ersten Eindruck von einer Funktion $f\colon D \to \mathbb{R}$ auf einer Teilmenge $D \subseteq \mathbb{R}$ zu gewinnen, startet man am besten mit einem Bild. Dieses kann zwar, falls D kein endliches Intervall ist, meist nur einen endlichen Ausschnitt des Definitionsbereichs zeigen und ist auch in seiner Genauigkeit begrenzt, bietet aber doch Information, die man anders nur mit erheblich größerem Aufwand darstellen könnte. Als Bild von f betrachten wir üblicherweise den Graphen

$$\Gamma_f \underset{\text{def}}{=} \left\{(x,y) \in \mathbb{R}^2 \;\middle|\; y = f(x)\right\},$$

also eine Teilmenge der Ebenen $\mathbb{R}^2$, die wir uns sehr gut vorstellen können.

Auch für eine Funktion $f\colon D \to \mathbb{R}^m$ mit $D \subseteq \mathbb{R}^n$ können wir deren Graphen

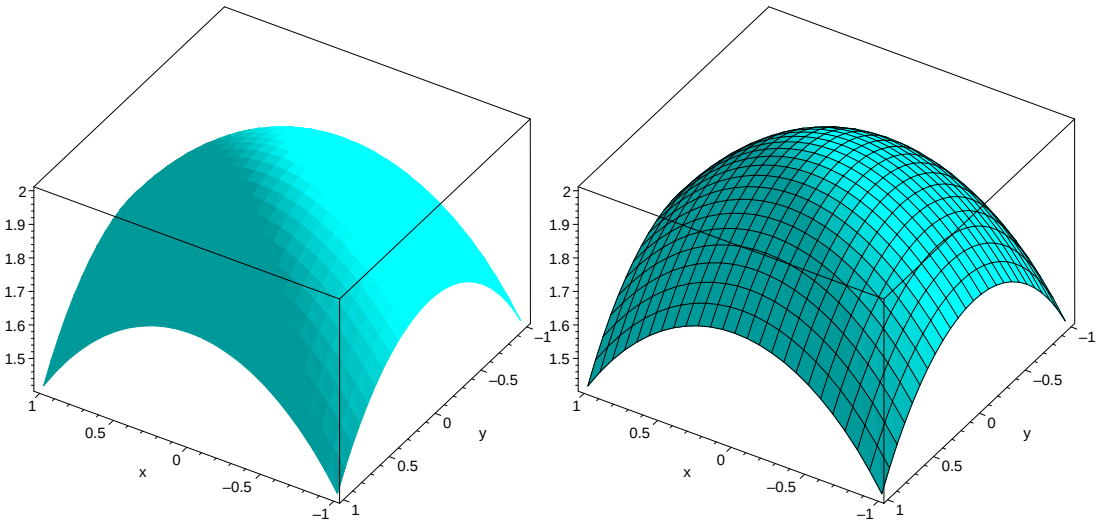
$$\Gamma_f \underset{\text{def}}{=} \left\{(x,y) \in D \times \mathbb{R}^m \;\middle|\; y = f(x)\right\}$$

definieren; da dieser in $\mathbb{R}^n \times \mathbb{R}^m = \mathbb{R}^{n+m}$ liegt, ist er allerdings nur für $n + m \leq 3$ wirklich anschaulich, wobei es im Fall $n + m = 3$ bei komplizierteren Funktionen stark von der gewählten Perspektive abhängen kann, wieviel man wirklich sieht. Für einfache reellwertige Funktionen zweier Veränderlicher jedoch ist der Graph sicherlich die beste Methode zur Veranschaulichung, wobei man gegebenenfalls zur besseren Übersicht noch Hilfslinien für die Funktionswerte zu ausgewählten Werten von x und/oder y einzeichnen kann.

Beim Graphen der Funktion

$$f\colon \begin{cases} [-1,1] \times [-1,1] & \to \mathbb{R} \\ (x,y) & \mapsto \sqrt{4 - x^2 - y^2} \end{cases}$$

in der Abbildung auf der nächsten Seite etwa sieht man bei beiden Darstellungen recht gut, daß Γ_f Teil einer Kugeloberfläche ist.



Graph der Funktion $f(x,y) = \sqrt{4 - x^2 - y^2}$

Eine andere Möglichkeit zur Veranschaulichung von Funktionen zweier Veränderlicher ist von topographischen Karten her bekannt: Dort wird die Höhe über dem Meeresspiegel, eine Funktion der beiden Ebenenkoordinaten, dargestellt durch *Höhenlinien.* Entsprechend können wir für eine beliebige Funktion $f\colon D \to \mathbb{R}$ mit $D \subseteq \mathbb{R}^2$ und jeden Wert $c \in \mathbb{R}$ die *Niveaulinie*

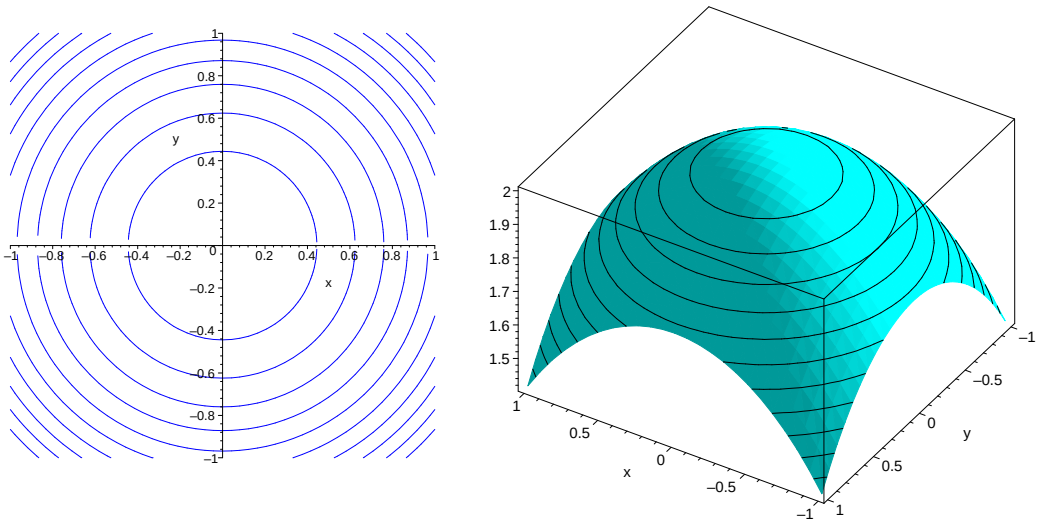
$$N_c(f) \underset{\text{def}}{=} \{(x,y) \in \mathbb{R}^2 \mid f(x,y) = c\}$$

definieren; sie muß natürlich keine „Linie“ sein, sondern kann auch nur aus einigen Punkten bestehen, leer sein oder – im Falle einer konstanten Funktion – für einen bestimmten Wert c aus dem gesamten Definitionsbereich D bestehen.

Im Falle des obigen Beispiels etwa ist $N_c(f)$ für $c > 2$ und für $c < \sqrt{2}$ die leere Menge; für $c = 2$ besteht sie nur aus dem Nullpunkt, und für $c = \sqrt{2}$ aus den vier Punkten $(0, \pm 1)$ und $(\pm 1, 0)$. Für $\sqrt{2} < c < 2$ erhalten wir die in der nächsten Abbildung für $c = 1{,}5$ bis $c = 2$ in Schritten von 0,05 dargestellten Kreislinien

$$\sqrt{4 - x^2 - y^2} = c \quad \text{oder} \quad x^2 + y^2 = 4 - c^2\,,$$

eingeschränkt natürlich auf das Einheitsquadrat als dem Definitionsbereich von f. Die Darstellung von Niveaulinien kann auch kombiniert werden mit der des Graphen.

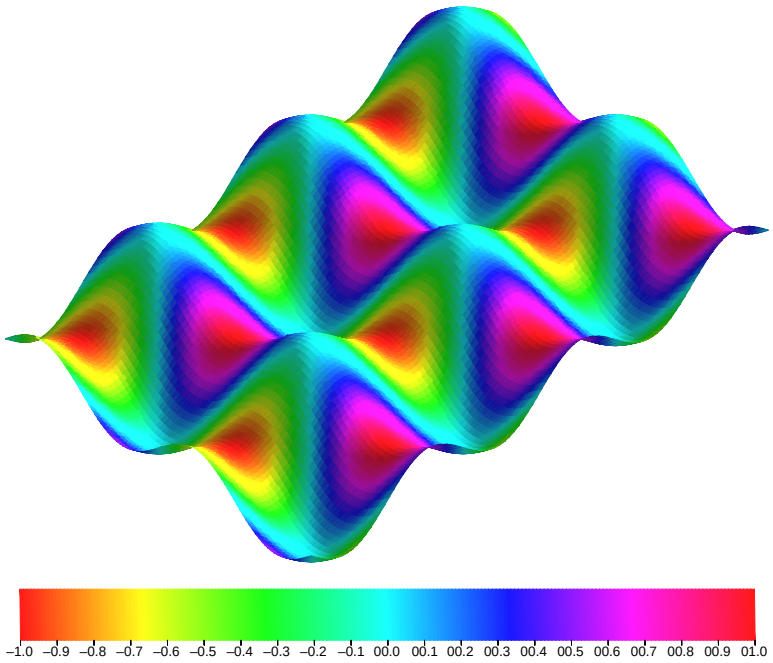


Niveaulinien der Funktion $f(x,y) = \sqrt{4 - x^2 - y^2}$

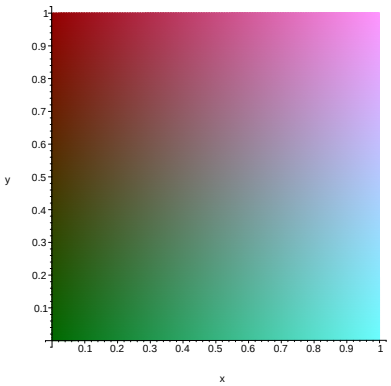
Für Funktionen von mehr als zwei Veränderlichen ist die Visualisierung naturgemäß schwieriger; wir können Graphen und auch Niveaulinien, -flächen *usw.* zwar problemlos definieren, aber nicht mehr zeichnen – es sei denn, es handelt sich um sehr einfache Niveauflächen im $\mathbb{R}^3$. Bei Funktionen mit Werten in einem mehrdimensionalen Raum kommt hinzu, daß die Niveaumengen dann nicht mehr nur von einem, sondern von mehreren Parametern abhängen.

Eine weitere Möglichkeit besteht darin, auf einem zwei- oder dreidimensionalen Graphen durch Farbe, Textur *usw.* weitere Dimensionen darzustellen; allgemein bekannt ist die Kodierung der Höhe durch von Grün nach Braun laufende Farben in Atlanten oder auch die Darstellung der Temperatur durch Farbverläufe von Blau über Rot nach Weiß, *usw.*

Wollen wir beispielsweise für $0 \le x \le 3\pi$ und $0 \le y \le 2\pi$ jene Funktion von $\mathbb{R}^2$ nach $\mathbb{R}^2$ veranschaulichen, die einem Punkt (x,y) die beiden Werte $f(x,y) = \sin(x+y)\cos(x-y)$ und $g(x,y) = \cos(x+y)\sin(x-y)$ zuordnet, so können wir den Graphen von f zeichnen und darauf nach einem Farbschema die Funktion g kodieren. Da diese nur Werte zwischen -1 und 1 annimmt, müssen wir dazu einfach ein Funktion festlegen, die jeder dieser Zahlen eine Farbe zuordnet; der entsprechende Farbverlauf ist unter dem Bild abgedruckt. Wenn man genau hinschaut, gewinnt man so einen recht guten Eindruck vom relativen Verlauf der beiden Funktionen.

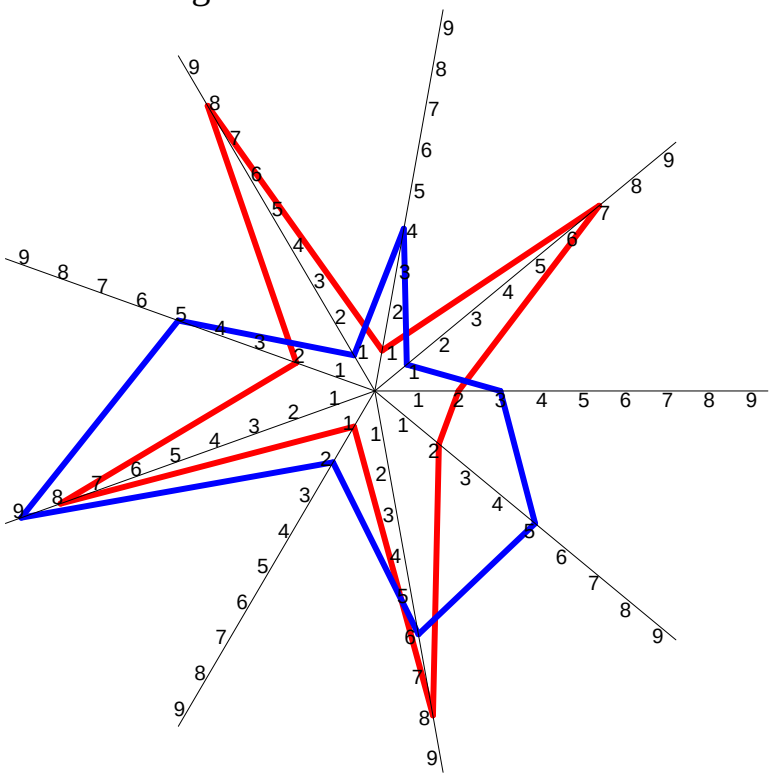


Grundsätzlich kann man mit Farben auch mehr als eine Dimension darstellen: Da wir drei Arten von Sehzäpfchen haben, könnten wir theoretisch bis zu drei Dimensionen darstellen. Tatsächlich sind jedoch nur die wenigsten Menschen in der Lage, einer Farbe deren Rot- Grün- und Blauwerte anzusehen, so daß die Darstellung von drei Dimensionen durch Farbwerte selten nützlich ist. Zwei Dimensionen sind aber durchaus realistisch, vor allem, wenn wir die eine Dimension auf die Helligkeit abbilden und die andere auf einen der beiden Farbwerte in einem Luminanz/ Chromanz-Modell wie etwa dem fürs jpeg-Format verwendeten YCbCr-Modell. Beim links abgebildeten Quadrat etwa sind die x-Werte durch die Helligkeit kodiert und die y-Werte durch die Chromanz für Rot.



Als letzte Alternative seien noch Stern- oder Netzdiagramme erwähnt: Hier geht es darum, einzelne Punkte in einem höherdimensionalen Raum zu veranschaulichen, z.B. den Vektor der Klausurnoten eines Studenten oder Kompetenzeinschätzungen für Politiker. Um einen Punkt $(x_1, \ldots, x_n) \in \mathbb{R}^n$ darzustellen, zeichnet man n vom Nullpunkt ausgehende Strahlen im $\mathbb{R}^2$, markiert auf dem i-ten dieser Strahlen den

Punkt x_i, und verbindet die so markierten Punkte zu einem n-Eck. Damit dies übersichtlich bleibt, sollte n hier nicht wesentlich größer als zehn sein, und auch die Anzahl der in einem Bild darstellbaren Punkte sollte definitiv einstellig sein.



Die Punkte $(3, 1, 4, 1, 5, 9, 2, 6, 5)$ und $2, 7, 1, 8, 2, 8, 1, 8, 2)$ aus $\mathbb{R}^9$

Ein eigenes Forschungsgebiet der Mathematik und Informatik, die Visualisierung, beschäftigt sich mit diesen und weiteren Methoden, die für eine vorgegebene Fragestellung interessanten Aspekte einer (analytisch oder empirisch gegebenen) Funktion mehrerer Veränderlichen graphisch herauszuarbeiten; für uns sollen aber zumindest vorerst Graphen und Niveaumengen genügen.

b) Normierte Vektorräume

Bei der Definition von Konvergenz und Stetigkeit im Eindimensionalen spielte die Betragsfunktion eine große Rolle; die in diesem Abschnitt eingeführten Normen sollen diese Funktion aufs Höherdimensionale verallgemeinern.

Genau wie die klassische Betragsfunktion jeder reellen (oder komplexen) Zahl eine nichtnegative reelle Zahl zuordnet, soll eine Norm jedem

Punkt des $\mathbb{R}^n$ (oder $\mathbb{C}^n$) eine nichtnegative reelle Zahl zuordnen. Wenn wir den Punkt $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ mit dem Vektor mit denselben Komponenten identifizieren, können wir beispielsweise dessen Länge

$$\sqrt{\langle x, x \rangle} = \sqrt{x_1^2 + \cdots + x_n^2}$$

nehmen, was oft in der Tat die natürlichste Wahl ist. Andererseits ist das Rechnen mit dieser Länge wegen der Wurzelfunktion gelegentlich recht unangenehm; deshalb wollen wir auch Alternativen betrachten.

Obwohl wir uns in diesem Semester praktisch nur für $\mathbb{R}^n$ interessieren, möchte ich zumindest die Definition für einen beliebigen reellen oder komplexen Vektorraum angeben:

Definition: Ein normierter Vektorraum ist ein $\mathbb{R}$- oder $\mathbb{C}$-Vektorraum V zusammen mit einer Abbildung $\|\cdot\| : V \to \mathbb{R}$ mit den Eigenschaften

a) $\|\lambda v\| = |\lambda| \, \|v\|$ für alle $\lambda \in \mathbb{R}$ *bzw.* $\mathbb{C}$ und $v \in V$

b) $\|v\| \geq 0$ für alle $v \in V$, und $\|\vec{v}\| = 0$ genau dann, wenn $v = 0$

c) $\|v + w\| \leq \|v\| + \|w\|$ *(Dreiecksungleichung)*

Die Zahl $\|v\|$ wird als *Norm* des Vektors $v \in V$ bezeichnet.

Im Falle $V = \mathbb{R}$ oder auch $V = \mathbb{C}$ ist klar, daß $\|v\| = |v|$ alle drei Forderungen erfüllt; der Begriff der Norm verallgemeinert also in der Tat den des Betrags.

Ist $\langle \cdot, \cdot \rangle$ ein Skalarprodukt auf dem reellen Vektorraum V, kann durch

$$\|v\| \underset{\text{def}}{=} \sqrt{\langle v, v \rangle}$$

eine Norm auf V definiert werden: Abgesehen von *c)* sind alle Forderungen aus der obigen Definition klar; zum Beweis von *c)* müssen wir beachten, daß wegen der Bilinearität eines Skalarprodukts gilt

$$\|v + w\|^2 = \langle v + w, v + w \rangle = \|v\|^2 + \|w\|^2 + 2 \langle v, w \rangle \qquad \text{und}$$

$$(\|v\| + \|w\|)^2 = \|v\|^2 + \|w\|^2 + 2 \|v\| \, \|w\| \ ,$$

so daß die Behauptung im Falle $\langle v, w \rangle \geq 0$ sofort aus der CAUCHY-SCHWARZschen Ungleichung folgt und für $\langle v, w \rangle < 0$ aus der Nichtnegativität der Norm.

Ausgehend vom Standardskalarprodukt

$$\langle v, w\rangle = \langle (v_1, \ldots, v_n), (w_1, \ldots, w_n)\rangle \underset{\text{def}}{=} v_1 w_1 + \cdots + v_n w_n$$

auf $\mathbb{R}^n$ erhalten wir so die EUKLIDische Norm

$$\|v\| = \|(v_1, \ldots, v_n)\| = \sqrt{v_1^2 + \cdots + v_n^2}\,,$$

mit der wir in dieser Vorlesung häufig arbeiten werden.

Der Hauptgrund dafür, daß wir auch noch andere Normen betrachten, liegt in der Unhandlichkeit des Umgangs mit der Wurzel. Die zweite für uns wichtige Norm auf $\mathbb{R}^n$, die *Maximumsnorm*, vermeidet dies. Sie ist definiert als

$$\|v\|_\infty = \|(v_1, \ldots, v_n)\|_\infty \underset{\text{def}}{=} \max\{|v_1|, \ldots, |v_n|\}\,.$$

Bedingung *a)* ist erfüllt, da $|\lambda v_i| = |\lambda| \cdot |v_i|$ für alle i, und auch mit *b)* gibt es keine Probleme. Für *c)* betrachten wir zwei Punkte

$$v = (v_1, \ldots, v_n) \quad \text{und} \quad w = (w_1, \ldots, w_n)$$

aus $\mathbb{R}^n$. Die Maximumsnorm der Summe

$$v + w = (v_1 + w_1, \ldots, v_n + w_n)$$

ist nach Definition der größte Wert eines Betrags $|v_i + w_i|$; dieser werde etwa für den Index j angenommen, d.h. $\|v + w\|_\infty = \big|v_j + w_j\big|$.

Die Norm $\|v\|_\infty$ von v ist der größte Betrag eines v_i und damit mindestens gleich $\big|v_j\big|$; genauso ist $\|w\|_\infty \geq \big|w_j\big|$. Also ist

$$\|v + w\|_\infty = \big|v_j + w_j\big| \leq \big|v_j\big| + \big|w_j\big| \leq \|v\|_\infty + \|w\|_\infty\,,$$

wie verlangt.

Damit ist auch die Maximumsnorm tatsächlich eine Norm. Sie kann allerdings nicht über ein Skalarprodukt auf $\mathbb{R}^n$ definiert werden: Gäbe es nämlich ein Skalarprodukt $\langle\cdot,\cdot\rangle$, für das $\|v\|_\infty = \sqrt{\langle v, v\rangle}$ wäre; so wäre etwa im $\mathbb{R}^2$ insbesondere

$$\left\langle \begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}1\\0\end{pmatrix} \right\rangle = \left\| \begin{pmatrix}1\\0\end{pmatrix} \right\|_\infty^2 = 1 \quad \text{und} \quad \left\langle \begin{pmatrix}0\\1\end{pmatrix}, \begin{pmatrix}0\\1\end{pmatrix} \right\rangle = \left\| \begin{pmatrix}0\\1\end{pmatrix} \right\|_\infty^2 = 1\,,$$

also

$$1 = \left\| \begin{pmatrix}1\\1\end{pmatrix} \right\|_\infty^2 = \left\langle \begin{pmatrix}1\\0\end{pmatrix} + \begin{pmatrix}0\\1\end{pmatrix}, \begin{pmatrix}1\\0\end{pmatrix} + \begin{pmatrix}0\\1\end{pmatrix} \right\rangle = 1+1+2 \left\langle \begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}0\\1\end{pmatrix} \right\rangle$$

und somit $\left\langle \begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}0\\1\end{pmatrix} \right\rangle = -\frac{1}{2}$ und $\left\langle \begin{pmatrix}2\\0\end{pmatrix}, \begin{pmatrix}0\\1\end{pmatrix} \right\rangle = -1$. Mithin wäre

$$4 = \left\| \begin{pmatrix}2\\1\end{pmatrix} \right\|_\infty^2 = \left\langle \begin{pmatrix}2\\0\end{pmatrix} + \begin{pmatrix}0\\1\end{pmatrix}, \begin{pmatrix}2\\0\end{pmatrix} + \begin{pmatrix}0\\1\end{pmatrix} \right\rangle = 4 + 1 - 2 = 3\,,$$

ein offensichtlicher Widerspruch.

Maximumsnormen lassen sich nicht nur für $\mathbb{R}^n$ (und analog $\mathbb{C}^n$) definieren, sondern auch für Funktionenräume: Wie wir aus Kapitel 2, §5 wissen, nimmt eine stetige Funktion $f\colon [a,\, b] \to \mathbb{R}$ auf einem *abgeschlossene* Intervall ihr Maximum wirklich an, d.h die Abbildung

$$\|\cdot\|_\infty : \mathcal{C}^0\big([a,\, b],\ \mathbb{R}\big) \to \mathbb{R}; \quad f \mapsto \max_{x\in[a,\, b]} |f(x)|$$

ist wohldefiniert. Sie hat auch die Eigenschaften *a)* bis *c): a)* und *b)* sind, wie in den meisten Fällen, trivial, und sind $f, g \in \mathcal{C}^0\big([a,\, b],\ \mathbb{R}\big)$ zwei Funktionen, so ist auch deren Summe $f + g$ stetig. Wenn sie ihr Betragsmaximum im Punkt $x^* \in [a,\, b]$ annimmt, haben wir die Abschätzung

$$\begin{aligned}\|f + g\|_\infty = |(f + g)(x^*)| = |f(x^*) + g(x^*)| &\le |f(x^*)| + |g(x^*)| \\ &\le \max_{x\in[a,\, b]} |f(x)| + \max_{x\in[a,\, b]} |g(x)| = \|f\|_\infty + \|g\|_\infty\ .\end{aligned}$$

Maximumsnormen spielen unter anderem in der Numerik eine wichtige Rolle, denn sie liefern Fehlerschranken für numerische Rechnungen:

Führen wir beispielsweise auf dem Vektorraum $\mathbb{R}^{n\times m}$ aller $n \times m$-Matrizen die Maximumsnorm ein, so setzen wir natürlich

$$\|A\|_\infty = \max_{i=1}^{n} \max_{j=1}^{m} \left|a_{ij}\right|\ .$$

Betrachten wir auch $\mathbb{R}^m$ und $\mathbb{R}^n$ mit der Maximumsnorm, so erhalten wir für einen Vektor $v \in \mathbb{R}^m$ und dessen Produkt $b = Av$ mit einer $n \times m$-Matrix A die Abschätzung

$$\|b\|_\infty \le m\,\|A\|_\infty \cdot \|v\|_\infty\ ,$$

denn nach der Multiplikationsregel für Matrizen ist $b_i = \sum_{j=1}^{m} a_{ij} v_j$ und damit

$$|b_i| \le \sum_{j=1}^{m} \left|a_{ij}\right| \cdot \left|v_j\right| \sum_{j=1}^{m} \|A\|_\infty \cdot \|v\|_\infty = m\,\|A\|_\infty\,\|v\|_\infty\ .$$

Wird also der Vektor v durch einen Fehlervektor ϵ gestört, so ist

$$A(v + \epsilon) = Av + A\epsilon = b + A\epsilon$$

mit einem Fehler behaftet, dessen Komponenten *mit Sicherheit* kleiner sind als $m \|A\|_\infty \cdot \|\epsilon\|_\infty$. Entsprechend ändert sich bei einem eindeutig lösbaren linearen Gleichungssystem $Ax = b$ die Lösung $x = A^{-1}b$ *höchstens* um $n \|A^{-1}\|_\infty \cdot \|\epsilon\|_\infty$, wenn die rechte Seite durch einen Vektor ϵ gestört wird. (Tatsächlich wird diese Schranke in den meisten Fällen viel zu pessimistisch sein, aber realistische Schranken sind in der Numerik oft – wenn überhaupt – nur mit sehr großem Aufwand zu finden.)

Wir haben Normen eingeführt als Verallgemeinerungen der Betragsfunktion, um damit auch Begriffe wie Konvergenz und Stetigkeit aufs Mehrdimensionale zu übertragen. Im Falle der Konvergenz führt dies zur

Definition: $\big(V, \|\cdot\|\big)$ sei ein normierter $\mathbb{R}$-Vektorraum. Eine Folge $v_1, v_2, \ldots$ von Vektoren aus V konvergiert gegen den Vektor $v \in V$, wenn es zu jedem $\varepsilon > 0$ eine natürliche Zahl $N \in \mathbb{N}$ gibt, so daß $\|v - v_n\| < \varepsilon$ für alle $n \geq N$.

Da es bei der Konvergenz nur darum geht, daß die Folgenglieder dem Grenzwert beliebig nahe kommen, ist klar, daß sich an der Konvergenz einer Folge nichts ändert, wenn wir die verwendete Norm durch ein positives Vielfaches ersetzen. Allgemeiner definieren wir

Definition: Zwei Normen $\|\cdot\|_1$ und $\|\cdot\|_2$ auf einen Vektorraum V heißen äquivalent, wenn es reelle Konstanten $c_1, c_2 > 0$ gibt, so daß

$$c_1 \|v\|_1 \leq \|v\|_2 \leq c_2 \|v\|_1 \ .$$

Offensichtlich ist dann auch

$$\frac{1}{c_2} \|v\|_2 \leq \|v\|_1 \leq \frac{1}{c_1} \|v\|_2 \ ,$$

die Äquivalenz ist also, wie es sein muß, symmetrisch.

Beispielsweise sind auf jedem $\mathbb{R}^n$ die EUKLIDische Norm $\|\cdot\|$ und die Maximumsnorm $\|\cdot\|_\infty$ äquivalent, denn

$$\|v\| = \sqrt{v_1^2 + \cdots + v_n^2} \leq \sqrt{n\,\|v\|_\infty^2} = \sqrt{n}\;\|v\|_\infty$$

und

$$\|v\|_\infty = \sqrt{\|v\|_\infty^2} \leq \sqrt{v_1^2 + \cdots + v_n^2} = \|v\|\ ,$$

d.h.

$$\|v\|_\infty \leq \|v\| \leq \sqrt{n}\;\|v\|_\infty\ .$$

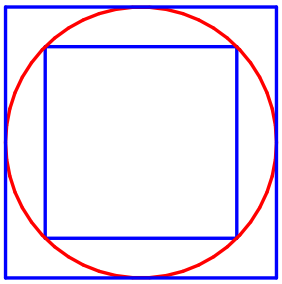
Anschaulich bedeutet dies, daß jeder Würfel in eine Kugel eingebettet werden kann und umgekehrt, denn

$$\left\{x \in \mathbb{R}^n \;\middle|\; \|x\|_\infty \leq a\right\}$$

ist ein Würfel mit Kantenlänge $2a$ und

$$\left\{x \in \mathbb{R}^n \;\middle|\; \|x\| \leq r\right\}$$

eine Kugel mit Radius r.



Für den Nachweis der Konvergenz einer Folge kann mal die eine, mal die andere dieser Normen besser geeignet sein. Zum Glück haben wir die freie Auswahl:

Lemma: $\|\cdot\|_1$ und $\|\cdot\|_2$ seien zwei äquivalente Normen auf dem Vektorraum V. Eine Folge $(v_n)_{n\in\mathbb{N}}$ von Vektoren aus V konvergiert genau dann bezüglich $\|\cdot\|_1$ gegen $v \in V$, wenn sie bezüglich $\|\cdot\|_2$ gegen v konvergiert.

Beweis: Da die Normen äquivalent sind, gibt es positive reelle Zahlen c_1, c_2, so daß $c_1\,\|v\|_1 \leq \|v\|_2 \leq c_2\,\|v\|_1$ ist für alle $v \in V$. Falls die Folge $(v_n)_{n\in\mathbb{N}}$ bezüglich $\|\cdot\|_1$ gegen $v \in V$ konvergiert, gibt es zu jedem $\varepsilon > 0$ ein $N \in \mathbb{N}$, so daß $\|v - v_n\|_1 < \varepsilon$ ist für alle $n \geq N$. Also gibt es bei vorgegebenem $\varepsilon > 0$ auch ein $M \in \mathbb{N}$, so daß $\|v - v_n\|_1 < \varepsilon/c_2$ für alle $n \geq M$. Für $n \geq M$ ist dann

$$\|v - v_n\|_2 \leq c_2\,\|v - v_n\|_1 \leq c_2\frac{\varepsilon}{c_2} = \varepsilon\,,$$

die Folge konvergiert also auch bezüglich $\|\cdot\|_2$ gegen v.

Da $\|v\|_1 \leq \|v\|_2 / c_1$ ist für alle $v \in V$, folgt ganz entsprechend, daß jede bezüglich $\|\cdot\|_2$ konvergente Folge auch bezüglich $\|\cdot\|_1$ gegen denselben Grenzwert konvergiert.

■

Man kann zeigen, daß in $\mathbb{R}^n$ alle Normen äquivalent sind, so daß es dort also nur einen Konvergenzbegriff gibt. Wir werden in $\mathbb{R}^n$ praktisch immer mit der EUKLIDischen Norm oder der Maximumsnorm arbeiten, deren Äquivalenz wir gezeigt haben; daher können wir immer, wenn von Konvergenz die Rede ist, frei wählen, mit welcher der beiden Normen wir arbeiten möchten. Je nach Anwendung kann

c) Stetigkeit

Stetigkeit bedeutet anschaulich, daß kleine Änderungen der Argumente einer Funktion auch nur zu kleinen Änderungen der Funktionswerte führen. Um Spielraum für die Variation der Argumente zu haben, definierten wir Stetigkeit für Funktionen einer Veränderlichen nur für Funktionen, die auf offenen Intervallen definiert sind. Im Falle mehrerer Veränderlicher wollen wir ähnlich vorgehen; wir brauchen daher als erstes eine mehrdimensionale Entsprechung für offene Intervalle.

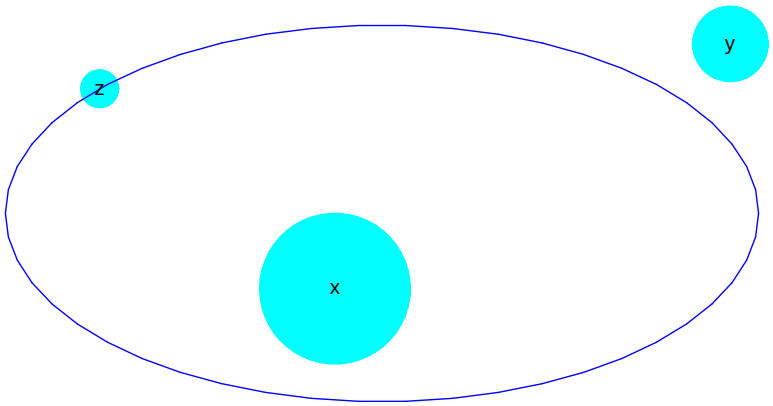
Die für uns wesentliche Eigenschaft eines offenen Intervalls $(a,\ b)$ war, daß es dort für jeden Punkt $x \in (a,\ b)$ sowohl links als auch rechts von x weitere Punkte aus dem Intervall gibt: Ist ε das Minimum der beiden (positiven) Werte $x - a$ und $b - x$, so liegt das Intervall $(x - \varepsilon,\ x + \varepsilon)$ ganz in $(a,\ b)$; wir können uns also sowohl nach links als auch nach rechts um einen beliebigen Betrag kleiner ε bewegen, ohne das Intervall zu verlassen.

Im $\mathbb{R}^2$ und erst recht in höherdimensionalen Räumen können wir uns nicht nur nach links und rechts bewegen, sondern in unendlich viele Richtungen; wir wollen verlangen, daß es wieder eine positive Zahl ε gibt, so daß wir uns in jeder Richtung um jeden Betrag echt kleiner ε bewegen können ohne die Menge zu verlassen. Im Zweidimensionalen bedeutet dies, wenn wir mit der EUKLIDischen Norm arbeiten, daß es zu jedem Punkt x eine Kreisscheibe um diesen Punkt geben soll, die ganz in der Menge drin liegt; in höheren Dimensionen haben wir entsprechend Kugeln und deren höherdimensionale Verallgemeinerungen.

Definition: *a)* Eine Teilmenge $M \subseteq V$ eines normierten Vektorraums V mit Norm $\|\cdot\|$ heißt *offen*, wenn es zu jedem Punkt $x \in M$ ein $\varepsilon > 0$ gibt, so daß jedes $y \in V$ mit $\|y - x\| < \varepsilon$ in M liegt.
b) M heißt *abgeschlossen*,, wenn die Komplementärmenge $V \setminus M$ offen ist.
c) $M \subseteq N$ heißt *offen in der Teilmenge* $N \subseteq V$, wenn es eine offene Teilmenge $U \subseteq V$ gibt, so daß $M = U \cap N$ ist; M heißt *abgeschlossen in* N, wenn es eine abgeschlossene Teilmenge $A \subseteq V$ gibt mit $M = A \cap N$.

Wir können dies auch anders formulieren, indem wir zunächst für eine beliebige Teilmenge $M \subseteq V$ und einen beliebigen Punkt $x \in V$ dessen Lage in Bezug auf M klassifizieren:

Definition: *a)* Ein Punkt $x \in V$ heißt *innerer Punkt* der Teilmenge $M \subseteq V$, wenn es ein $\varepsilon > 0$ gibt, so daß alle $y \in V$ mit $\|x - y\| < \varepsilon$ in M liegen.
b) x heißt *äußerer Punkt* von M, wenn es ein $\varepsilon > 0$ gibt, so daß alle $y \in V$ mit $\|x - y\| < \varepsilon$ *nicht* in M liegen.
c) x heißt *Randpunkt* von M, wenn es für jedes $\varepsilon > 0$ einen Punkt $y \in M$ gibt mit $\|x - y\| < \varepsilon$ sowie einen Punkt $z \in V \setminus M$, so daß $\|x - z\| < \varepsilon$.



x ist innerer, y äußerer Punkt und z Randpunkt der blau umrandeten Menge

Ein innerer Punkt x von M muß also insbesondere in der Menge M drin liegen, während ein äußerer Punkt von M *nicht* in M liegen darf. In beiden Fällen müssen auch noch alle hinreichend nahe bei x liegenden Punkte dieselbe Eigenschaft haben.

Ein Randpunkt muß nicht in der Menge liegen, kann es aber. Wichtig ist, daß es beliebig nahe von x sowohl Punkte gibt, die in der Menge liegen, als auch solche, die dies nicht tun.

Wenn wir im $\mathbb{R}^n$ mit der EUKLIDischen Norm arbeiten, ist ein Punkt $x \in M$ genau dann ein innerer Punkt von M, wenn es eine Kreisscheibe *bzw.* n-dimensionale Kugel um x gibt, die ganz in M liegt; ein Punkt x aus $\mathbb{R}^n \setminus M$ ist genau dann äußerer Punkt, wenn auch noch eine Kreisscheibe *bzw.* n-dimensionale Kugel um x ganz außerhalb von M liegt. Von einem Randpunkt schließlich verlangen wir, daß jede noch so kleine Kreisscheibe *bzw.* n-dimensionale Kugel um x sowohl Punkte aus M enthält als auch solche, die nicht in M liegen.

Wenn wir stattdessen mit der Maximumsnorm arbeiten, gilt fast dasselbe; nur haben wir jetzt an Stelle der Kreisscheiben Quadrate und an Stelle von n-dimensionalen Kugeln n-dimensionale Würfel. Die inneren, äußeren und Randpunkte sind offensichtlich bezüglich beider Normen dieselben, und wie oben im Falle der Konvergenz zeigt man leicht

Lemma: Sind $\|\cdot\|_1$ und $\|\cdot\|_2$ äquivalente Normen auf dem Vektorraum V, so ist ein Punkt $x \in V$ genau dann innerer, äußerer *bzw.* Randpunkt der Teilmenge $M \subseteq V$ bezüglich $\|\cdot\|_1$, wenn er diese Eigenschaft bezüglich $\|\cdot\|_2$ hat.

■

Ebenfalls ziemlich klar ist

Lemma: *a)* Eine Teilmenge $M \subseteq V$ eines normierten Vektorraums V ist genau dann offen, wenn jeder Punkt $x \in M$ ein innerer Punkt ist.
b) M ist genau dann abgeschlossen, wenn jeder Randpunkt von M in M liegt.

Beweis: a) folgt sofort aus dem Vergleich der Definition offener Mengen und innerer Punkte. Für *b)* betrachten wir zunächst eine abgeschlossene Teilmenge $M \subseteq V$ und einen Randpunkt x von M. Läge x nicht in M, müßte x in $V \setminus M$ liegen. Da $V \setminus M$ eine offene Menge ist, gäbe es also ein $\varepsilon > 0$, so daß auch alle $y \in V$ mit $\|x - y\| < \varepsilon$ in $V \setminus M$ liegen müßten. Dies widerspricht aber der Definition eines Randpunkt, für den es mindestens ein $y \in M$ geben muß mit $\|x - y\| < \varepsilon$.

Ist umgekehrt $M \subseteq V$ eine Menge, die jeden ihrer Randpunkte enthält, ist keiner der Punkte $x \in V \setminus M$ Randpunkt von M, es gibt also zu jedem $x \in V \setminus M$ ein $\varepsilon > 0$, so daß entweder alle $y \in V$ mit $\|x - y\| < \varepsilon$ in M liegen oder aber alle solche y in $V \setminus M$ liegen. Ersteres ist nicht möglich, da x selbst nicht in M liegt, obwohl $\|x - x\| = 0 < \varepsilon$ ist; daher müssen alle diese y in $V \setminus M$ liegen, so daß $V \setminus M$ eine offene Menge ist. Damit ist M selbst abgeschlossen, wie behauptet. ∎

Nach diesen Vorbereitungen können wir nun stetige Funktionen definieren:

Definition: *a)* Eine Abbildung $f\colon D \to W$ von einer Teilmenge $D \subseteq V$ eines normierten Vektorraums $\left(V, \|\cdot\|_1\right)$ in einen normierten Vektorraum $\left(W, \|\cdot\|_2\right)$ heißt *stetig* in $x \in D$, wenn es für jedes $\varepsilon > 0$ ein $\delta > 0$ gibt, so daß für alle $y \in D$ gilt: Ist $\|x - y\|_1 < \delta$, so ist $\|f(x) - f(y)\|_2 < \varepsilon$.
b) f heißt *stetig*, wenn f in jedem Punkt $x \in D$ stetig ist.

Auch hier überzeugt man sich leicht, daß sich nichts ändert, wenn wir $\|\cdot\|_1$ und/oder $\|\cdot\|_2$ durch eine äquivalente Norm ersetzen.

Konkret für eine Funktion von n reellen Variablen mit Werten in $\mathbb{R}$ besagt diese Definition, ausgedrückt für die Maximumsnorm

Definition: Eine Funktion $f\colon D \to \mathbb{R}$ auf einer Teilmenge D des $\mathbb{R}^n$ heißt *stetig* im Punkt $x = (x_1, \ldots, x_n) \in D$, wenn es zu jedem $\varepsilon > 0$ ein $\delta > 0$ gibt, so daß für jeden Punkt $y = (y_1, \ldots, y_n) \in D$ gilt: Falls $|y_i - x_i| < \delta$ ist für alle i, dann ist $|f(y) - f(x)| < \varepsilon$.

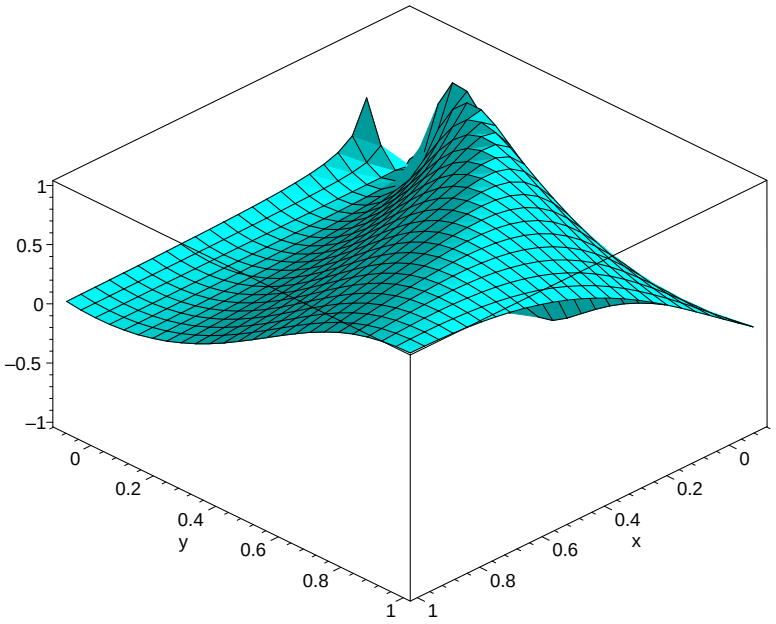
Im Eindimensionalen kann man der graphischen Darstellung einer Funktion leicht ansehen, ob sie stetig ist oder nicht; wir wollen schauen, wie dies im Mehrdimensionalen aussieht.

Der Einfachheit halber beschränken wir uns auf Funktionen zweier Veränderlicher, z.B. die Funktion

$$f\colon \begin{cases} \mathbb{R}^2 & \to \mathbb{R} \\ (x,y) & \mapsto \begin{cases} \dfrac{2xy^2}{x^2+y^4} & \text{falls } (x,y) \neq (0,0) \\ 0 & \text{falls } (x,y) = (0,0) \end{cases} \end{cases} .$$

Wir sollten erwarten (und werden auch gleich sehen), daß diese Funktion auf $\mathbb{R}^2 \setminus \{(0,0)\}$ stetig ist, denn dort ist sie nur über Grundrechenarten definiert, wobei Division durch Null ausgeschlossen ist, da $x^2 + y^4$ nur im Nullpunkt verschwindet. Bleibt also der Punkt $(0,0)$ zu untersuchen.

Die Abbildung unten zeigt den Graphen von f in einer kleinen Umgebung des Nullpunkts. Sie zeigt zwar einen relativ steilen Sprung entlang der Geraden $x = 0$, aber nur für die etwas weiter vom Nullpunkt entfernten x-Werte; bei $y = 0$ sieht alles harmlos und ziemlich eben aus.



Ist diese Funktion stetig?

Nun ist der abgebildete Graph natürlich von einem Computer anhand von nur endlich vielen Stützpunkten konstruiert; um wirklich zu entscheiden, ob f im Nullpunkt stetig ist, können wir uns nicht auf diese Approximation verlassen, sondern müssen die Funktion etwas genauer untersuchen.

Da wir uns im Eindimensionalen recht gut auskennen, können wir beispielsweise die Einschränkungen von f auf die verschiedenen Geraden durch den Nullpunkt betrachten. Abgesehen von der y-Achse haben diese alle die Form $y = ax$ mit $a \in \mathbb{R}$, und

$$f(x, ax) = \frac{2x \cdot (ax)^2}{x^2 + (ax)^4} = \frac{2a^2x^3}{x^2(1 + a^4x^2)} = \frac{2a^2x}{1 + a^4x^2}$$

für $x \neq 0$. Für $x \to 0$ geht beim rechtsstehenden Ausdruck der Zähler gegen Null und der Nenner gegen eins; der Grenzwert existiert also und

ist gleich $f(0,0) = 0$, d.h. die Einschränkung von f ist stetig auf der Geraden $y = ax$.

Auf der y-Achse verschwindet f für jeden Wert von x, ist also ebenfalls stetig, d.h. die Einschränkung von f auf jede Gerade durch den Nullpunkt ist stetig.

Betrachten wir zur Vorsicht auch noch die Einschränkung von f auf die Parabel $x = ay^2$! Dort ist

$$f(ay^2, y) = \frac{2ay^2 \cdot y^2}{a^2y^4 + y^4} = \frac{2ay^4}{(1 + a^2)y^4} = \frac{2a}{1 + a^2}$$

für alle $y \neq 0$, wohingegen $f(0,0) = 0$ ist. Für $a \neq 0$ ist die Einschränkung von f auf diese Parabel also nicht stetig, und damit kann auch f nicht stetig sein: Setzen wir $a = 1$, so ist $2a/(1 + a^2) = 1$, die Funktion nimmt also beliebig nahe beim Nullpunkt den Wert eins an. Formal können wir die Unstetigkeit bezüglich der Maximumsnorm folgendermaßen beweisen:

Wenn f im Nullpunkt stetig wäre, gäbe es zu jedem $\varepsilon > 0$, ein $\delta > 0$, so daß für alle Punkte (x, y) mit $\|(x, y)\|_\infty = \max\{|x|, |y|\} < \delta$ der Betrag von $f(x, y)$ kleiner als ε wäre. Für jedes $\delta > 0$ und jedes y mit $|y| < \delta$ und $|y| < 1$ ist aber $\left|y^2\right| < |y| < \delta$, also $\left\|(y^2, y)\right\|_\infty < \delta$, aber

$$f(y^2, y) = \frac{2y^2 \cdot y^2}{y^4 + y^4} = 1\,,$$

so daß es für $\varepsilon = \frac{1}{2}$ kein solches δ geben kann.

Dem Graphen konnten wir das nicht ansehen, was bei näherer Überlegung eigentlich auch nicht verwundert: Der Graph wurde von einem Computer gezeichnet, und der kann natürlich nur eine endliche Auswahl $x_1, \ldots, x_n$ *bzw.* $y_1, \ldots, y_m$ von Werten berücksichtigen. Dazu berechnet er die Funktionswerte $f(x_i, y_j)$ und verbindet dann die Punkte zu gleichem x_i *bzw.* gleichem y_j durch Kurven. Entlang dieser Kurven ist f aber, wie wir gesehen haben, stetig.

Anders sieht es aus, wenn wir stattdessen die Niveaulinien von f betrachten: Zusammen mit den Koordinatenachsen überdecken die Parabeln $x = ay^2$ mit $a \in \mathbb{R} \setminus \{0\}$ den gesamten $\mathbb{R}^2$, die Niveaulinie $N_0(f)$

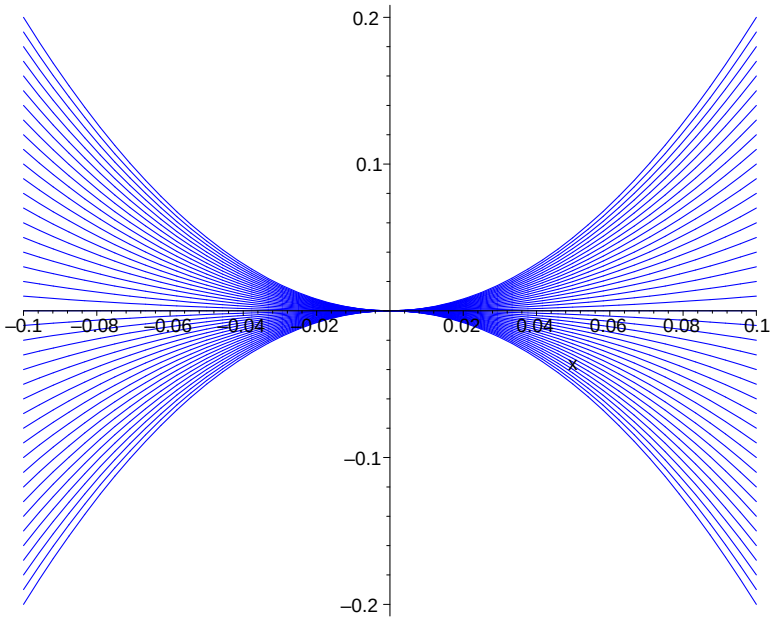
von f besteht also aus den beiden Koordinatenachsen, während die anderen Niveaulinien aus Parabeln $x = ay^2$ jeweils ohne den Nullpunkt bestehen; siehe Abbildung. Da die Gleichung

$$\frac{2a}{1+a^2} = c$$

für $c \neq 0$ die beiden Lösungen

$$a = \frac{1 \pm \sqrt{1-c^2}}{c}$$

hat, besteht jede Niveaulinie für $0 < |c| < 1$ aus zwei dieser Parabeln, für $|c| = 1$ aus einer, und für $|c| > 1$ ist $N_c(f) = \emptyset$.



Die Niveaulinien illustrieren die Unstetigkeit im Nullpunkt

Dies zeigt auch anschaulich, daß f im Nullpunkt unstetig ist, denn alle Niveaulinien kommen dem Nullpunkt beliebig nahe, obwohl dieser nur auf $N_0(f)$ liegt. Daran sehen wir übrigens auch, daß der oben abgebildete Graph falsch ist: In jeder Umgebung von $(0,0)$ wird jeder Wert c zwischen -1 und 1 angenommen, der Abschluß des Graphen enthält also die Strecke $[-1,1]$ auf der z-Achse. Wegen der oben beschriebenen Vorgehensweise von Maple (und auch fast aller anderer Computergraphikprogramme) ist das aber in der Abbildung nicht zu sehen.

Bei Funktionen, in deren Definition Fallunterscheidungen eingehen, ist also größere Vorsicht geboten als im Eindimensionalen; bei in der Pra-

xis auftretenden Funktionen wird es allerdings wohl meist so sein, daß die Unstetigkeitsstellen genau dort auftreten, wo man Sprünge definiert hat. Schwierig wird es nur, wenn man wie im obigen Beispiel in einem Punkt eine Situation der Art „$0/0$“ hat und entscheiden muß, ob man stetig ergänzen kann: Hier hilft im Mehrdimensionalen keine DE L'HOSPITALsche Regel, es hilft auch nicht, die Annäherung der Funktion an den problematischen Punkt aus allen Richtungen zu untersuchen, sondern man muß wirklich auf die Definition der Stetigkeit zurückgehen.

Zum Glück sind Funktionen wie die obige allerdings nicht der Regelfall, mit dem wir es in Anwendungen zu tun haben; für die meisten gängigen Funktionen ist wie im Eindimensionalen ziemlich klar, daß sie stetig sind.

Beginnen wir mit linearen Funktionen! Da ein konstanter Summand an der Form des Graphen nicht ändert, wollen dabei im Gegensatz zur Linearen Algebra auch inhomogene Funktionen betrachten:

Lemma: Jede lineare Funktion

$$f\colon \begin{cases} \mathbb{R}^n \to \mathbb{R} \\ x = (x_1, \ldots, x_n) \mapsto \sum_{j=1}^{n} a_j x_j + b \end{cases}$$

ist stetig.

Beweis: $x = (x_1, \ldots, x_n)$ sei ein fester Punkt aus $\mathbb{R}^n$ und $\varepsilon > 0$ sei eine positive reelle Zahl. Für einen weiteren Punkt $u = (u_1, \ldots, u_n) \in \mathbb{R}^n$ ist dann

$$\begin{aligned} |f(u) - f(x)| = \left| \sum_{j=1}^{n} a_j (u_j - x_j) \right| &\leq \sum_{j=1}^{n} |a_j| \cdot |u_j - x_j| \\ &\leq \sum_{j=1}^{n} |a_j| \cdot \|u - x\|_\infty \ . \end{aligned}$$

Falls alle a_j verschwinden, ist dies gleich Null, also insbesondere kleiner

als ε; andernfalls ist es kleiner als ε, falls

$$\|u - x\|_\infty < \delta = \frac{\varepsilon}{\sum_{j=1}^{n} |a_j|} .$$

In beiden Fällen folgt, daß f stetig im Punkt x ist und damit stetig auf ganz $\mathbb{R}^n$, denn $x \in \mathbb{R}^n$ war ein beliebiger Punkte.

∎

Damit haben wir zwar nur lineare Funktionen mit Werten in $\mathbb{R}$, aber das reicht: Eine Funktion $f\colon D \to \mathbb{R}^m$ ordnet jedes $x \in D$ einen Punkt $f(x) \in \mathbb{R}^m$ zu, und dieser ist gegeben durch m reelle Zahlen $f_1(x), \ldots, f_m(x)$. Bezeichnen wir die so definierten Funktionen $f_i\colon D \to \mathbb{R}$ als die *Komponenten* von f, so gilt:

Lemma: Eine Funktion $f\colon D \to \mathbb{R}^m$ auf einer Teilmenge $D \subseteq \mathbb{R}^n$ ist genau dann stetig in einem Punkt $x \in D$, wenn jede ihrer Komponenten $f_i\colon D \to \mathbb{R}$ stetig in x ist.

Beweis: Wir arbeiten mit den Maximumsnormen auf $\mathbb{R}^n$ und $\mathbb{R}^m$. Falls f in $x \in D$ stetig ist, gibt es zu jedem $\varepsilon > 0$ ein $\delta > 0$, so daß

$$\|f(u) - f(x)\|_\infty = \max\bigl\{ |f_i(u) - f_i(x)| \;\big|\; i = 1, \ldots, n \bigr\} < \varepsilon$$

falls $\|u - x\|_\infty < \delta$. In diesen Fall ist erst recht $|f_i(u) - f_i(x)| < \varepsilon$ für alle i, also sind alle f_i stetig in x.

Umgekehrt seien alle f_i stetig in x. Dann gibt es für alle $\varepsilon > 0$ positive reelle Zahlen $\delta_1, \ldots, \delta_n > 0$, so daß

$$|f_i(u) - f_i(x)| < \varepsilon \quad \text{falls} \quad \|u - x\|_\infty < \delta_i .$$

Bezeichnet δ die kleinste unter den n Zahlen δ_i, ist daher

$$\|f(u) - f(x)\|_\infty = \max\bigl\{ |f_i(u) - f_i(x)| \;\big|\; i = 1, \ldots, n \bigr\} < \varepsilon$$

falls $\|u - x\|_\infty < \delta$. Somit ist auch f stetig in x.

∎

Da jede Komponente einer linearen Funktion linear ist, folgt daraus sofort

Lemma: Jede lineare Funktion

$$f\colon \begin{cases} \mathbb{R}^n \to \mathbb{R}^m \\ x = (x_1, \ldots, x_n) \mapsto \bigl(f_1(x), \ldots, f_m(x)\bigr) \end{cases}$$

mit $f_i(x) = \sum_{j=1}^{n} a_{ij}x_j + b_i$ ist stetig.

■

Die folgende Charakterisierung stetiger Funktionen ist gelegentlich einfacher anzuwenden als die Definition:

Lemma: Eine Funktion $f\colon D \to \mathbb{R}^m$ auf einer Teilmenge $D \subseteq \mathbb{R}^n$ ist genau dann stetig, wenn für jede offene Menge $U \subseteq \mathbb{R}^m$ das Urbild $f^{-1}(U) = \{x \in D \mid f(x) \in U\}$ offen in D ist.

Beweis: Sei zunächst f stetig auf D und $U \subseteq \mathbb{R}^m$ eine offene Menge. Wir müssen zeigen, daß $f^{-1}(U)$ eine offene Menge ist.

Wir betrachten einen festen Punkt $x \in D$ mit $f(x) \in U$. Wegen der Offenheit von U gibt es dann ein $\varepsilon > 0$, so daß alle $y \in \mathbb{R}^m$ mit $\|y - f(x)\| < \varepsilon$ in U liegen. Zu diesem ε wiederum gibt es wegen der Stetigkeit von f in x in $\delta > 0$, so daß für alle $u \in D$ mit $\|x - u\| < \delta$ gilt: $\|f(x) - f(u)\| < \varepsilon$. Setzen wir

$$W_x = \{u \in \mathbb{R}^n \mid \|x - u\| < \delta\},$$

so ist also $f(u) \in U$ für alle $u \in W_x \cap D$.

Die Vereinigung W alle W_x mit $f(x) \in U$ ist eine offene Menge, denn jeder Punkt von W liegt in einer der Mengen W_x und ist wegen der Offenheit von W_x dort und somit erst recht in W ein innerer Punkt. Damit ist $f^{-1}(U) = W \cap D$ offen in D.

Umgekehrt habe f die Eigenschaft, daß das Urbild einer jeden offenen Menge offen in D ist. Wir müssen zeigen, daß f in jedem Punkt $x \in D$ stetig ist.

Dazu sei $\varepsilon > 0$ und $U = \{y \in \mathbb{R}^m \mid \|f(x) - y\| < \varepsilon\}$. Dies ist eine offene Menge, also ist ihr Urbild $f^{-1}(U)$ offen in D. Es gibt daher eine offene Menge $W \subseteq \mathbb{R}^n$, so daß $f^{-1}(U) = W \cap D$ ist. Da x in $f^{-1}(U)$ liegt, ist x ein innerer Punkt von W; es gibt also ein $\delta > 0$, so daß alle $u \in \mathbb{R}^n$ mit $\|u - x\| < \delta$ in W liegen. Jedes $u \in D$ mit $\|u - x\| < \delta$ liegt daher in $W \cap D = f^{-1}(U)$. Somit ist $\|f(x) - f(u)\| < \varepsilon$ für alle $u \in D$ mit $\|u - x\| < \delta$. Dies zeigt die Stetigkeit von f in x.

■

Lemma: $f\colon D \to \mathbb{R}^m$ und $g\colon E \to \mathbb{R}^p$ seien stetige Funktionen auf den Teilmengen $D \subseteq \mathbb{R}^n$ und $E \subseteq \mathbb{R}^m$, und $f(D)$ liege in E. Dann ist auch die Hintereinanderausführung

$$h = g \circ f\colon \begin{cases} D \to \mathbb{R}^p \\ x \mapsto g\big(f(x)\big) \end{cases}$$

stetig.

Beweis: Dies folgt am einfachsten aus dem vorigen Lemma: Ist $U \subseteq \mathbb{R}^p$ eine offene Menge, so ist wegen der Stetigkeit von g auch $g^{-1}(U) \subseteq \mathbb{R}^m$ offen in E, d.h. es gibt eine offene Teilmenge $W \subset \mathbb{R}^m$, so daß $g^{-1}(U) = W \cap E$ ist. Da Punkte, die nicht in E liegen, kein Urbild unter f haben, ist $f^{-1}\big(g^{-1}(U)\big) = f^{-1}(W)$, was wegen der Stetigkeit von f offen in D ist. $f^{-1}\big(g^{-1}(U)\big)$ ist aber gerade das Urbild von U unter der zusammengesetzten Abbildung $g \circ f$.

∎

Lemma: *a)* Die Funktion $f\colon \mathbb{R}^2 \to \mathbb{R}$ mit $f(x,y) = x \pm y$ ist stetig.
b) Die Funktion $g\colon \mathbb{R}^2 \to \mathbb{R}$ mit $g(x,y) = xy$ ist stetig.
c) Die Funktion $h\colon \mathbb{R} \times \big(\mathbb{R} \setminus \{0\}\big) \to \mathbb{R}$ mit $h(x,y) = x/y$ ist stetig.

Beweis: a) ist klar, da es sich hier um lineare Funktionen handelt, und deren Stetigkeit haben wir bereits allgemein bewiesen.

Der Beweis von *b)* beruht im wesentlichen auf demselben Trick, mit dem wir in Kapitel 2 bewiesen haben, daß für zwei konvergente Folgen die Produktfolge gegen das Produkt aus deren Grenzwerten konvergiert. Wie dort müssen wir die Fälle, daß einer oder gar beide Faktoren verschwinden, gesondert behandelt.

Sei also zunächst $(x,y) = (0,0)$ und $\varepsilon > 0$. Dann ist auch $\delta = \sqrt{\varepsilon} > 0$ und für alle Punkte $(u,v) \in \mathbb{R}^2$ mit $\|(u,v)\|_\infty < \delta$ sind $|u|$ und $|v|$ kleiner als δ, also ist

$$|uv - xy| = |uv| = |u| \cdot |v| < \delta^2 = \varepsilon\,.$$

Damit ist die Stetigkeit im Punkt $(0,0)$ gezeigt.

Nun sei $x \neq 0$, aber $y = 0$. Für $(u,v) \in \mathbb{R}^2$ ist dann

$$|uv - xy| = |uv - 0| = |u(v-y)| = |u| \cdot |v-y|\ .$$

Wir müssen zu einem vorgegebenen $\varepsilon > 0$ ein $\delta > 0$ finden, so daß dies kleiner ist als ε falls

$$\|(x,y)-(u,v)\|_\infty = \max\{|x-u|, |y-v|\} = \max\{|x-u|, |v|\} < \delta .$$

Wählen wir $\delta \leq \frac{1}{2}|x|$, so erfüllt jedes u mit $|x-u| < \delta$ die Ungleichung

$$|u| < |x| + \frac{1}{2}|x| = \frac{3}{2}|x| .$$

Wählen wir δ so, daß zusätzlich auch noch $\delta \leq 2\varepsilon/3|x|$ ist, also beispielsweise

$$\delta = \min\left\{\frac{|x|}{2}, \frac{\varepsilon}{2|x|}\right\} ,$$

so ist

$$|uv-xy| = |u|\cdot|v-y| < \frac{3|x|}{2}\cdot\frac{2\varepsilon}{3|x|} = \varepsilon$$

für alle $(u,v) \in \mathbb{R}^2$ mit $\|(x,y)-(u,v)\|_\infty < \delta$.

Bleibt noch der Fall $y \neq 0$. Hier haben wir für einen Punkt $(u,v) \in \mathbb{R}^2$ die Abschätzung

$$|uv-xy| = |v(u-x)+x(v-y)| \leq |v|\cdot|u-x| + |x|\cdot|v-y| .$$

Falls wir nur $v \in \mathbb{R}$ mit $|v-y| < \frac{1}{2}|y|$ betrachten, können wir $|v|$ durch $\frac{3}{2}|y|$ nach oben abschätzen und erhalten die Ungleichung

$$|uv-xy| \leq \frac{3|y|}{2}\cdot|u-x| + |x|\cdot|v-y| .$$

Ist nun ein $\varepsilon > 0$ vorgegeben, setzen wir

$$\delta = \begin{cases} \frac{2\varepsilon}{3|y|} & \text{falls } x = 0 \\ \min\left\{\frac{\varepsilon}{3|y|}, \frac{\varepsilon}{2|x|}\right\} & \text{falls } x \neq 0 \end{cases} ;$$

dann ist $|uv-xy| < \varepsilon$ für alle $(u,v) \in \mathbb{R}^2$ mit $\|(u,v)-(x,y)\|_\infty < \delta$. Somit ist die Funktion stetig in jedem Punkt $(x,y) \in \mathbb{R}^2$.

Die in *c)* behauptete Stetigkeit der Division könnten wir ähnlich beweisen: schneller geht es aber, wenn wir die aus der Analysis I bekannte Stetigkeit der Funktion $\mathbb{R}\setminus\{0\} \to \mathbb{R}$ mit $y \mapsto 1/y$ ausnutzen. Da eine

Funktion genau dann stetig ist, wenn alle ihre Komponenten stetig sind, ist auch die Funktion

$$\begin{cases} \mathbb{R} \times \big(\mathbb{R} \setminus \{0\}\big) \to \mathbb{R} \times \big(\mathbb{R} \setminus \{0\}\big) \\ (x, y) \mapsto \left(x, \dfrac{1}{y}\right) \end{cases}$$

stetig, nach dem vorigen Lemma also auch ihre Schachtelung mit der Multiplikation, die Division. Damit haben wir die Stetigkeit aller Grundrechenarten bewiesen. ∎

Die gerade bewiesenen Lemmata lassen sich zusammenfassen zum folgenden Prinzip:

Wenn eine Funktion $f\colon D \to \mathbb{R}^m$ auf einer offenen Teilmenge $D \subseteq \mathbb{R}^n$ so aus Grundrechenarten und stetigen Funktionen zusammengesetzt ist, daß nie eine Funktion außerhalb ihrer Stetigkeitsbereichs verwendet wird und auch Divisionen durch Null ausgeschlossen sind, dann ist f stetig auf D.

§2: Differenzierbare Funktionen

Nachdem wir wissen, was Konvergenz und Stetigkeit im Mehrdimensionalen bedeuten, können wir uns der Differenzierbarkeit zuwenden. Wir beginnen mit einer kurzen Wiederholung des eindimensionalen Falls:

a) Funktionen einer Veränderlichen

Wir bezeichnen eine Funktion $f\colon (a, b) \to \mathbb{R}$ als differenzierbar im Punkt $x \in (a, b)$, wenn sie in dessen Umgebung durch eine lineare Funktion angenähert werden kann. In Kapitel 3 hatten wir dies so formuliert, daß es ein $c \in \mathbb{R}$ sowie eine Funktion $\widetilde{f}$ mit $\widetilde{f}(0) = 0$ geben muß, so daß für kleine Werte von h gilt

$$f(x + h) = f(x) + ch + h\widetilde{f}(h)\,.$$

Da diese Formulierung mit explizit angegebener Fehlerfunktion $\widetilde{f}$ bei längeren Betrachtungen recht umständlich ist, wollen wir eine abkürzende Sprechweise einführen, die LANDAUsche o-Notation: Wir schreiben $o(h)$, sobald wir *irgendeine* uns nicht weiter interessierende

Funktion von h haben, die für $h \to 0$ schneller gegen Null geht als h selbst, d.h.

$$\varphi(h) = o(h) \iff \lim_{h\to 0} \frac{\varphi(h)}{h} = 0\,.$$

$o(h)$ ist hier also keine Funktion, sondern steht für eine ganze Klasse von Funktionen; beispielsweise ist

$$h^2 = o(h), \quad h^5 = o(h) \quad \text{und} \quad h \cdot \sin h = o(h)\,,$$

aber $\sin h$ können wir nicht als $o(h)$ schreiben, denn

$$\lim_{h\to 0} \frac{\sin h}{h} = 1\,.$$

Entsprechend schreiben wir auch

$$\varphi(h) = o\big(\psi(h)\big), \quad \text{wenn} \quad \lim_{h\to 0} \frac{\varphi(h)}{\psi(h)} = 0$$

ist.

EDMUND GEORG HERMANN LANDAU (1877–1938) wurde in Berlin geboren und studierte an der dortigen Universität, wo er auch von 1899 bis 1909 lehrte. Dann bekam er einen Ruf an die damals führende deutsche Mathematikfakultät in Göttingen. 1933 verlor er seinen dortigen Lehrstuhl, denn die Studenten boykottierten seine Vorlesungen, da sie meinten, sie könnten Mathematik nur von einem Professor ihrer eigenen Rasse lernen. LANDAUS zahlreiche Publikationen beschäftigen sich vor allem mit der Zahlentheorie, über die er auch ein bedeutendes Lehrbuch schrieb; sehr bekannt sind seine Arbeiten über die Verteilung von Primzahlen.

Mit LANDAUS o-Notation können wir kurz sagen, die Funktion f sei genau dann differenzierbar in x mit Ableitung $f'(x)$, wenn

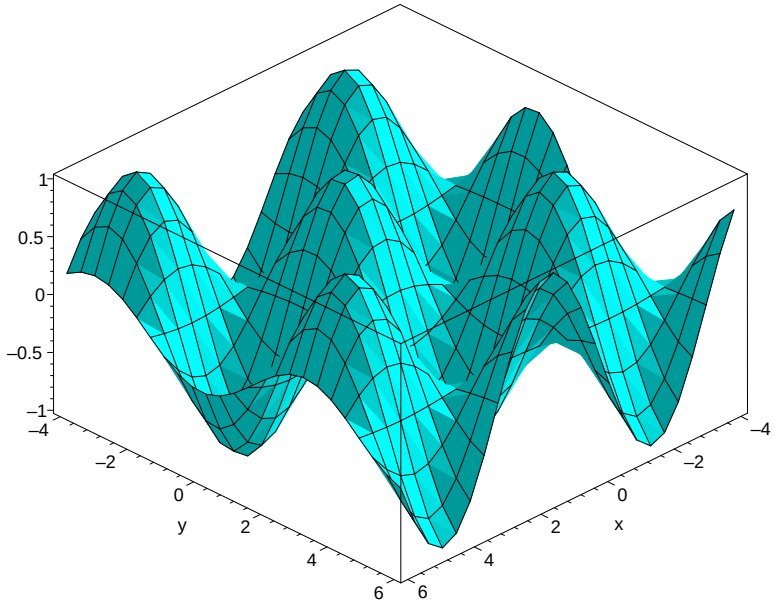
$$f(x + h) = f(x) + hf'(x) + o(h)$$

ist, denn natürlich ist $h\widetilde{f}(h) = o(h)$, denn der Quotient $h\widetilde{f}(h)/h = \widetilde{f}(h)$ geht für $h \to 0$ gegen $\widetilde{f}(0) = 0$.

b) Differenzierbarkeit im Mehrdimensionalen

Um Differenzierbarkeit für Funktionen mehrerer Veränderlicher zu definieren, können wir ähnlich vorgehen wie im eindimensionalen Fall.

Wir betrachten zunächst eine Funktion $f\colon \mathbb{R}^2 \to \mathbb{R}$, zum Beispiel $f(x,y) = \sin x \cos y$ in der Umgebung des Punktes $(1,1)$. Indem wir ihren Graphen sukzessive um den Faktor fünf vergrößern, erhalten wir die unten folgende Abbildungen. Sie zeigen, daß sich der Graph in einer hinreichend kleinen Umgebung von $(1,1)$ nur wenig von einer Ebenen unterscheidet, d.h. die Funktion ist dort annähernd linear.



Graph der Funktion $z = \sin x \cos y$

Eine lineare Funktion zweier Veränderlicher, bei der wir auch hier wieder im Gegensatz zur Linearen Algebra einen konstanten Term zulassen, läßt sich schreiben als

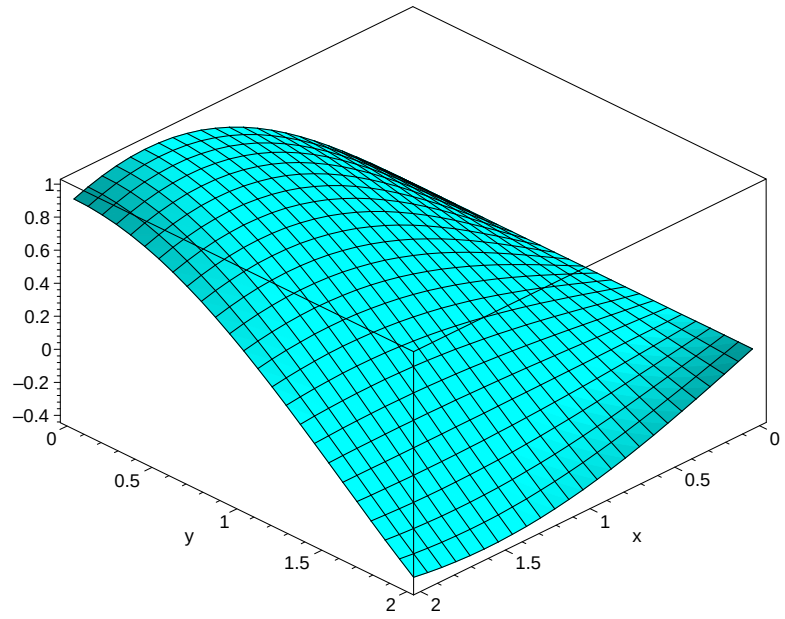
$$L(x,y) = a + bx + cy;$$

also ist

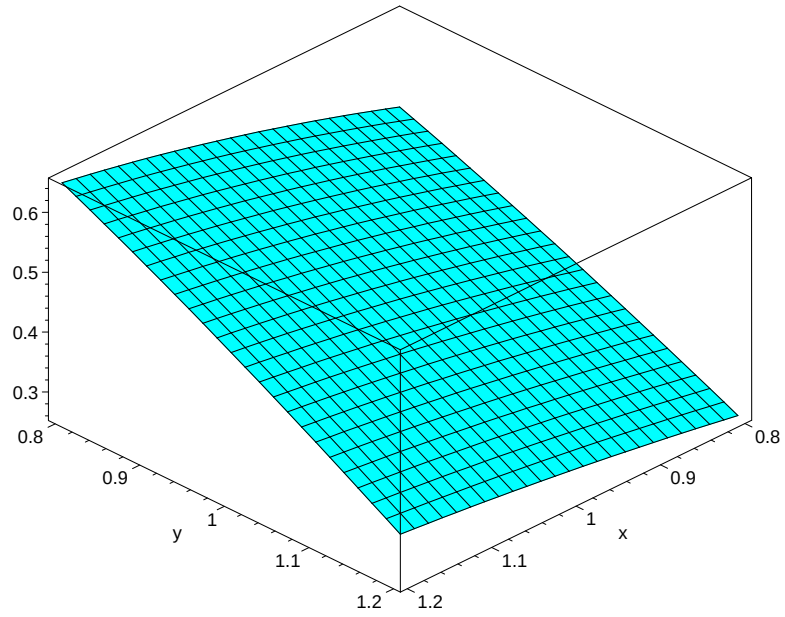
$$L(x+h, y+k) = a + b(x+h) + c(y+k) = L(x,y) + bh + ck$$

eine Approximation für f in der Umgebung des betrachteten Punkts (x,y). Im Punkt (x,y) sollte $L(x,y)$ natürlich mit $f(x,y)$ übereinstimmen, so daß $a = f(x,y)$ sein muß, und für $(h,k) \neq (0,0)$ sollte der Unterschied zwischen f und L schneller gegen Null gehen als der Abstand zwischen $(x+h, y+k)$ und (x,y), d.h. schneller als $\sqrt{h^2+k^2}$. Wir erwarten daher, daß

$$f(x+h, y+k) = f(x,y) + bh + ck + o(\sqrt{h^2+k^2})$$



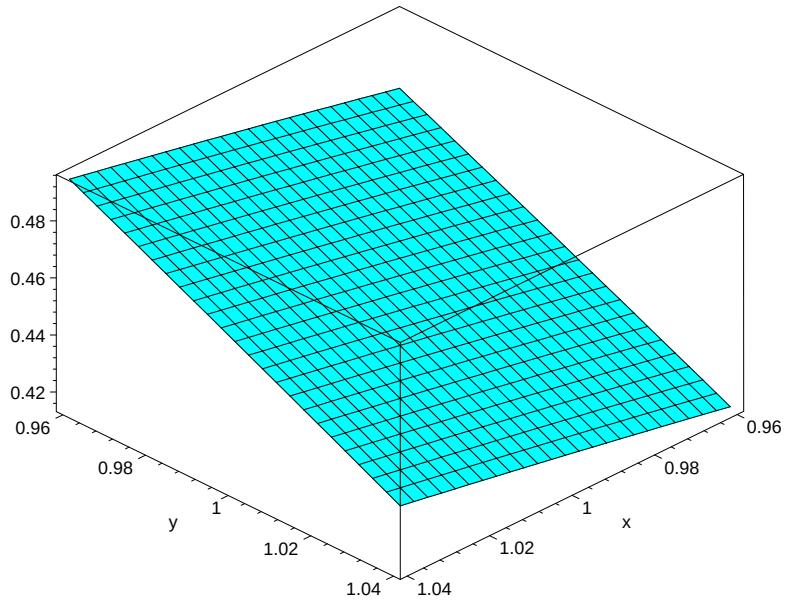
Vergrößerung um den Faktor fünf um $\big(1, 1, \sin(1)\cos(1)\big)$



Nochmalige Vergrößerung um den Faktor fünf um $\big(1, 1, \sin(1)\cos(1)\big)$

ist, und genau so läßt sich Differenzierbarkeit allgemein definieren.

Wenn wir eine Funktion von n Veränderlichen in der Umgebung eines Punkts $x \in \mathbb{R}^n$ betrachten, müssen wir alle n Variablen variieren; die Nachbarpunkte sind also Punkte der Form $x + h$ mit Vektoren $h \in \mathbb{R}^n$. Im Falle einer Funktion mit Werten in $\mathbb{R}$ ist $f(x + h)$ eine reelle Zahl



Nach noch einer Vergrößerung sieht die Funktion praktisch linear aus

und eine Linearisierung von f um x hat die Form

$$f(x) + \sum_{i=1}^{n} \gamma_i h_i \,,$$

wobei h_i die Komponenten des Vektors h sind. Für eine differenzierbare Funktion erwarten wir, daß der Fehler dieser Linearisierung schneller gegen Null geht als der Vektor h, das heißt also, schneller als dessen Norm. Ob wir dabei die EUKLIDische oder die Maximumsnorm verwenden, bleibt sich gleich, denn die beiden sind ja äquivalent. Wir verlangen daher einfach, daß der Fehler $o(\|h\|)$ sein soll, ohne uns auf eine spezielle Norm festzulegen.

Für eine Funktion mit Werten in $\mathbb{R}^m$ ändert sich nichts wesentliches: $f\colon D \to \mathbb{R}^m$ ist gegeben durch m Komponentenfunktionen $f_i\colon D \to \mathbb{R}$, und eine Linearisierung von f ist gleichbedeutend mit der Linearisierung der sämtlichen f_i. Im Falle der Differenzierbarkeit soll es also für jedes i reelle Zahlen $\gamma_{i1}, \ldots, \gamma_{in}$ geben, so daß

$$f_i(x + h) = f_i(x) + \sum_{j=1}^{n} \gamma_{ij} h_j + o(\|h\|)$$

ist. Den Vektor aus $\mathbb{R}^m$, dessen i-te Komponente die Summe der $\gamma_{ij} h_j$

ist, können wir auch kurz schreiben als Produkt

$$\begin{pmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1n} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{m1} & \gamma_{m2} & \cdots & \gamma_{mn} \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \\ \vdots \\ h_n \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{n} \gamma_{1j} h_j \\ \sum_{j=1}^{n} \gamma_{2j} h_j \\ \vdots \\ \sum_{j=1}^{n} \gamma_{mj} h_j \end{pmatrix}$$

der Matrix mit Einträgen γ_{ij} mit dem Vektor h. Für eine differenzierbare Funktion $f\colon D \to \mathbb{R}^m$ erwarten wir, daß es in jeder der m Komponenten einen Fehler der Form $o(\|h\|)$ gibt; wir haben also einen *Vektor* aus $\mathbb{R}^m$, dessen sämtliche Komponenten $o(\|h\|)$ sind. Um eine kurze Schreibweise zu haben, bezeichnen wir auch diesen Vektor einfach mit $o(\|h\|)$.

Wie im Eindimensionalen wollen wir Differenzierbarkeit in einem Punkt so definieren, daß die Funktion in einer Umgebung dieses Punktes durch eine lineare Funktion angenähert werden kann; wie dort müssen wir dazu sicherstellen, daß es genügend Punkte in der Umgebung gibt.

Definition: Ein Punkt $x_0 \in \mathbb{R}^n$ heißt *Häufungspunkt* von $D \subseteq \mathbb{R}^n$, wenn es für jedes $\varepsilon > 0$ unendlich viele Punkte $x \in D$ gibt mit $\|x - x_0\| < \varepsilon$.

Selbstverständlich ist jeder innere Punkt einer Menge $D \subseteq \mathbb{R}^n$ Häufungspunkt, denn dann liegen für hinreichend kleine Werte von ε ja sogar alle Punkte aus $\mathbb{R}^n$ mit $\|x - x_0\| < \varepsilon$ in D, aber zusätzlich sind beispielsweise auch die Ecke eine abgeschlossenen Quadrats D Häufungspunkte von D, auch wenn dann nur die Punkte aus einem Kreissegment mit Radius ε in D liegen.

Definition: *a)* Die Funktion $f\colon D \to \mathbb{R}^m$ mit $D \subseteq \mathbb{R}^n$ heißt *differenzierbar* im Punkt $x \in D$, wenn x ein Häufungspunkt von D ist und es eine $m \times n$-Matrix $J_f(x)$ gibt, so daß für Vektoren $h \in \mathbb{R}^n$ mit $h \to 0$ und $x + h \in D$ gilt

$$f(x + h) = f(x) + J_f(x)h + o(\|h\|)\,.$$

Die Matrix $J_f(x)$ heißt *Ableitung* oder JACOBI-Matrix von f im Punkt x.
b) Für eine Funktion $f\colon \mathbb{R}^n \to \mathbb{R}$ kann die dann einzeilige JACOBI-Matrix

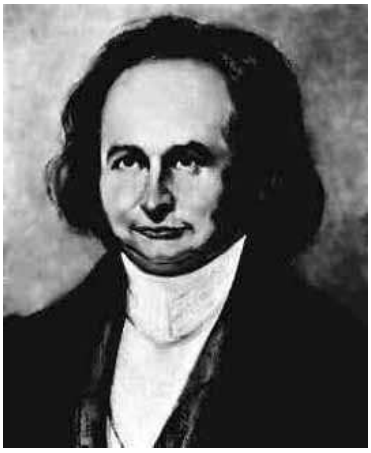
mit einem Vektor aus $\mathbb{R}^n$ identifiziert werden; dieser Vektor

$$\operatorname{grad} f(x) \underset{\text{def}}{=} \nabla f(x) \underset{\text{def}}{=} \begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_n \end{pmatrix}$$

heißt *Gradient* von f im Punkt $x \in D$.
c) Eine differenzierbare Funktion heißt *stetig differenzierbar,* wenn ihre Ableitung stetig ist.

∇ sieht zwar aus wie ein griechischer Buchstabe, ist aber keiner; es ist ein auf den Kopf gestelltes großes Delta (Δ). ∇f wird „Nabla f“ ausgesprochen nach dem griechischen Wort $\nu\alpha\beta\lambda\alpha$ = Leier; die Bezeichnung wurde eingeführt von dem irischen Mathematiker WILLIAM ROWEN HAMILTON, den die Form von ∇ an eine Leier erinnerte.

WILLIAM ROWEN HAMILTON (1805–1865) wurde in Dublin geboren; bereits mit fünf Jahren sprach er Latein, Griechisch und Hebräisch. Mit dreizehn begann er, mathematische Literatur zu lesen, mit 21 wurde er, noch als Student, Professor der Astronomie am Trinity College in Dublin. Er verlor allerdings schon bald sein Interesse für Astronomie und arbeitete weiterhin auf dem Gebiet der Mathematik und Physik. Am bekanntesten ist seine Entdeckung der Quaternionen 1843, vorher publizierte er aber auch bedeutende Arbeiten über Optik, Dynamik und Algebra.



CARL GUSTAV JACOB JACOBI (1804–1851) wurde in Potsdam als Sohn eines jüdischen Bankiers geboren und erhielt den Vornamen Jacques Simon. Im Alter von zwölf Jahren bestand er sein Abitur, mußte aber noch vier Jahre in Abschlußklasse des Gymnasiums bleiben, da die Berliner Universität nur Studenten mit mindestens 16 Jahren aufnahm. 1824 beendete er seine Studien mit dem Staatsexamen für Mathematik, Griechisch und Latein und wurde Lehrer. Außerdem promovierte er 1825 und begann mit seiner Habilitation. Etwa gleichzeitig konvertierte er zum Christentum, so daß er ab 1825 an der Universität Berlin und ab 1826 in Königsberg lehren konnte. 1832 wurde er dort Professor. Zehn Jahre später mußte er aus gesundheitlichen Gründen das raue Klima Königsbergs verlassen und lebte zunächst in Italien, danach für den Rest seines Lebens in Berlin. Er ist vor allem berühmt durch seine Arbeiten zur Zahlentheorie und über elliptische Integrale.

Mit obiger Definition haben wir etwas ähnliches wie die Definition

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

für Funktionen einer Veränderlicher: Auf diese Weise ist die Ableitung zwar *definiert*, aber sie wird – außerhalb von Anfängervorlesungen – praktisch nie so ausgerechnet. Entsprechend sollten wir auch für Funktionen mehrerer Veränderlicher eine effizientere Methode der Differentiation finden.

Am einfachsten wäre es, wenn wir den wohlbekannten Kalkül der Differentialrechnung für Funktionen einer Veränderlicher benutzen könnten, also versuchen wir, aus einer Funktion $f\colon D \to \mathbb{R}^m$ mit $D \subseteq \mathbb{R}^n$ Funktionen einer Veränderlichen zu machen.

Dabei können wir uns sofort auf den Fall $m = 1$ beschränken, denn jede Funktion $f\colon D \to \mathbb{R}^m$ ist zusammengesetzt aus m Komponentenfunktionen $f_i\colon D \to \mathbb{R}$; wir beschränken uns daher zunächst auf Funktionen $f\colon D \to \mathbb{R}$.

Schwieriger ist die Reduktion von n auf eins. Trotz der schlechten Erfahrungen im vorigen Paragraphen, wo eine unstetige Funktion nach Einschränkung auf eine beliebige Gerade stets stetig wurde, wollen wir uns exakt diese Einschränkungen genauer anschauen.

Eine Gerade durch einen gegebenen Punkt x ist eindeutig festgelegt durch einen Richtungsvektor e, wobei umgekehrt der Vektor e durch die Gerade natürlich *nicht* eindeutig festgelegt ist: Jedes Vielfache von e (außer dem Nullvektor) definiert dieselbe Gerade.

Wenn wir die Einschränkung von f auf eine solche Gerade mit Richtungsvektor e betrachten, betrachten wir konkret die Funktion

$$g(t) = f(x + te)\,,$$

einer einzigen Variablen $t \in \mathbb{R}$, die überall dort definiert ist, wo $x + te$ im Definitionsbereich D von f liegt; für eine offene Menge D also zumindest in einem gewissen offenen Intervall um den Nullpunkt der reellen Geraden.

Damit können wir nach der Differenzierbarkeit dieser Funktion für $t = 0$

fragen; falls sie differenzierbar ist, bezeichnen wir die Ableitung

$$g'(0) = \lim_{h\to 0} \frac{g(h) - g(0)}{h} = \lim_{h\to 0} \frac{f(x + he) - f(x)}{h}$$

als *Richtungsableitung* von f in Richtung e. Eine einfache Anwendung der Kettenregel, die jeder Leser am Rand des Skriptums kurz durchführen sollte, zeigt, daß diese „Richtungsableitung“ nicht nur von der *Richtung* des Vektors e abhängt, sondern auch von dessen *Länge:* Beispielsweise ist für $k(t) = f(x + 2te)$

$$k'(0) = 2g'(0)\,.$$

Speziell können wir diese Richtungsableitungen betrachten für den Fall, daß e ein *Einheitsvektor* ist (genau ist der Grund für die Bezeichnung e), beispielsweise einer der Koordinateneinheitsvektoren

$$e_i = (0, \ldots, 1, \ldots, 0)\,,$$

bei dem an der i-ten Stelle eine Eins steht und sonst lauter Nullen.

Alsdann ist für $g_i(t) = f(x + te_i)$

$$\begin{aligned} g_i'(0) &= \lim_{h\to 0} \frac{f(x + he_i) - f(x)}{h} \\ &= \lim_{h\to 0} \frac{f(x_1, \ldots, x_i + h, \ldots, x_n) - f(x_1, \ldots, x_i, \ldots, x_n)}{h} \end{aligned}$$

die Ableitung jener Funktion, die nur von x_i abhängt, während alle anderen Koordinaten x_j festgehalten werden. Wir behandeln also beim Differenzieren alle Variablen x_j mit Ausnahme von x_i als Konstanten und leiten die so entstehende Funktion in der üblichen Weise ab nach x_i. Diese Ableitung, so sie existiert, bezeichnen wir als *partielle Ableitung*

$$f_{x_i}(x) = \frac{\partial f}{\partial x_i}(x)$$

von f nach x_i; das Symbol ∂ wird, wenn überhaupt, als „del“ ausgesprochen, wobei *del* natürlich eine Abkürzung für *delta* ist. Partielle Ableitungen, so sie existieren, lassen sich nach den üblichen Regeln der Differentialrechnung für Funktionen einer Veränderlichen berechnen, sind also für „gutartige“ Funktionen problemlos.

Falls die Funktion f in $x \in D$ differenzierbar ist, existiert auch jede

Richtungsableitung, denn da dann für jeden Vektor h gilt

$$f\big(x+h\big) = f(x) + \langle \nabla f(x), h\rangle + o(\|h\|)\,,$$

ist insbesondere auch

$$f(x+te) = f(x) + \langle \nabla f(x), te\rangle + o(\|te\|) = f(x) + t\,\langle \nabla f(x), e\rangle + o(t)\,;$$

denn $\langle \nabla f(x), e\rangle$ und $\|e\|$ sind schließlich Konstanten. Damit existiert

$$\frac{d}{dt} f\big(x+te\big)_{|t=0} = \langle \nabla f(x), e\rangle$$

für jeden Richtungsvektor e; insbesondere existieren natürlich alle partiellen Ableitungen.

Die Umkehrung gilt leider nicht immer: Die (offensichtlich in $(0,0)$ unstetige) Funktion

$$f\colon \begin{cases} \mathbb{R}^2 & \to \mathbb{R} \\ (x,y) & \mapsto \begin{cases} 1 & \text{falls } xy \neq 0 \\ 0 & \text{falls } xy = 0 \end{cases} \end{cases}$$

verschwindet auf der x-Achse und der y-Achse; überall sonst hat sie den Wert eins. Damit existieren im Punkt $(0,0)$ beide partielle Ableitungen und sind identisch Null. Trotzdem ist f natürlich nicht differenzierbar in $(0,0)$, denn da $f(h,k)$ für jeden Punkt, der nicht auf einer der beiden Koordinatenachsen liegt, gleich eins ist, kann es keine Zahlen $b, c \in \mathbb{R}$ geben, so daß

$$f(h,k) = f(0,0) + bh + ck + o\big(\sqrt{h^2+k^2}\big)$$

ist, denn $f(0,0) = 0$ und $f(h,k) = 1$ für $hk \neq 0$.

Nun wird natürlich jeder vernünftige Mensch einwenden, daß dieses Beispiel sehr künstlich ist, und in der Tat verhalten sich „gutartige“ Funktionen nicht so:

Lemma: Falls $f\colon D \to \mathbb{R}$ auf der offenen Teilmenge $D \subseteq \mathbb{R}^n$ stetig ist und auch alle partiellen Ableitungen von f dort existieren und stetig sind, ist f in D differenzierbar und

$$\operatorname{grad} f(x) = \nabla f(x) = \begin{pmatrix} f_{x_1}(x) \\ \vdots \\ f_{x_n}(x) \end{pmatrix}\,.$$

Beweis: Sei $x \in D$. Da D offen ist, gibt es eine Kugel um x, die ganz in D liegt; wir betrachten im folgenden nur Vektoren h, deren Länge höchstens gleich dem Radius dieser Kugel ist, so daß $x + h$ stets in D liegt.

Wir betrachten f zunächst nur als Funktion der ersten Variablen; da deren Ableitung f_{x_1} in ganz D existiert, ist

$$f(x+h) = f(x_1 + h_1, \ldots, x_n + h_n)$$
$$= f(x_1, x_2 + h_2, \ldots, x_n + h_n) + f_{x_1}(x_1, x_2 + h_2, \ldots, x_n + h_n)h_1 + o(h_1)\,.$$

Genauso ist, da die partielle Ableitung nach x_2 in ganz D existiert,

$$f(x_1, x_2 + h_2, \ldots, x_n + h_n) =$$
$$f(x_1, x_2, x_3, \ldots, x_n + h_n) + f_{x_2}(x_1, x_2, x_3 + h_3 \ldots, x_n + h_n)h_2 + o(h_2)$$

und so weiter. Insgesamt erhalten wir

$$\begin{aligned} f(x+h) = f(x) &+ f_{x_1}(x_1, x_2 + h_2, x_3 + h_3, \ldots x_n + h_n)h_1 + o(h_1) \\ &+ f_{x_2}(x_1, x_2, x_3 + h_3, \ldots, x_n + h_n)h_2 + o(h_2) \\ &\vdots \\ &+ f_{x_{n-1}}(x_1, \ldots, x_{n-1}, x_n + h_n)h_{n-1} + o(h_{n-1}) \\ &+ f_{x_n}(x_1, \ldots, x_n)h_n + o(h_n)\,. \end{aligned}$$

Die LANDAU-Symbole $o(h_1), \ldots, o(h_n)$ können wir zu $o(\|h\|)$ zusammenfassen, denn da $|h_i| \leq \|h\|$ für jedes i, kann keines der h_i langsamer gegen Null gehen als $\|h\|$.

Damit sind wir schon ziemlich nahe an dem, was wir für die Differenzierbarkeit brauchen; allerdings hängen die partiellen Ableitungen noch von den h_i ab, so daß die Differenz zwischen $f(x + h)$ und $f(x)$ nicht durch eine lineare Funktion angenähert ist.

Hier kommt nun die Stetigkeit der partiellen Ableitungen ins Spiel: Diese impliziert, daß

$$\lim_{h \to 0} \big(f_{x_1}(x_1, x_2 + h_2, \ldots, x_n + h_n) - f_{x_1}(x_1, x_2, \ldots, x_n)\big) = 0$$

ist. Damit ist

$$\big(f_{x_1}(x_1, x_2 + h_2, \ldots, x_n + h_n) - f_{x_1}(x_1, x_2, \ldots, x_n)\big)h_1 = o(\|h\|)\,,$$

denn wenn h_1 mit einem Ausdruck multipliziert wird, der gegen Null geht, strebt das Produkt für $h \to 0$ schneller gegen Null als h_1 allein, und $o(h_1)$ kann durch $o(\|h\|)$ abgeschätzt werden. Also ist

$$\begin{aligned} & f_{x_1}(x_1, x_2 + h_2, \ldots, x_n + h_n)h_1 \\ =& f_{x_1}(x_1, x_2, \ldots, x_n)h_1 + o(\|h\|) = f_{x_1}(x)h_1 + o(\|h\|)\,. \end{aligned}$$

Entsprechend können wir auch bei den übrigen partiellen Ableitungen argumentieren und erhalten insgesamt, daß

$$\begin{aligned} f(x+h) &= f(x) + f_{x_1}(x)h_1 + \cdots + f_{x_n}(x)h_n + o(\|h\|) \\ &= f(x) + \begin{pmatrix} f_{x_1} \\ \vdots \\ f_{x_n} \end{pmatrix} \cdot h + o(\|h\|) \end{aligned}$$

ist. Damit ist das Lemma bewiesen. ∎

Für Funktionen mit stetigen partiellen Ableitungen ist der Gradient also gerade der Vektor der partiellen Ableitungen; er kann damit über die bekannten Ableitungsregeln für Funktionen einer Veränderlichen berechnet werden.

NB: Häufig wird der Gradient durch diese Formel *definiert;* in diesem Fall folgt natürlich aus der Existenz des Gradienten nicht die Differenzierbarkeit der Funktion; siehe obiges Beispiel einer unstetigen Funktion, für die alle partiellen Ableitungen in $(0, 0)$ existieren.

c) Ableitungsregeln

Da wir Ableitungen von Funktionen mehrerer Veränderlichen auf die mit Methoden der eindimensionalen Analysis bestimmbaren partiellen Ableitungen zurückgeführt haben, sollten wir erwarten, daß sich auch die gewohnten Rechenregeln der Differentialrechnung einer Veränderlichen übertragen lassen. Dies ist in der Tat der Fall, wobei selbst die Beweismethoden fast wörtlich übernommen werden können. Daher sollen hier nur ganz kurz einige einfache Regeln gezeigt werden; für den Rest sei auf die Übungen verwiesen.

Am einfachsten ist die Linearität der Ableitung:

Lemma: Sind $f, g\colon D \to \mathbb{R}^m$ differenzierbare Funktionen auf $D \subseteq \mathbb{R}^n$, so ist auch für alle $a, b \in \mathbb{R}$ die Funktion $af + bg$ differenzierbar und $J_{af+bg}(x) = aJ_f(x) + bJ_g(x)$ für alle $x \in D$. Speziell im Falle $m = 1$ ist also $\nabla(af + bg)(x) = a\nabla f(x) + b\nabla g(x)$.

Beweis: Wegen der Differenzierbarkeit von f und g ist für alle $x \in D$ und $h \in \mathbb{R}^n$ mit $x + h \in D$

$$f(x+h) = f(x)+J_f(x)h+o(\|h\|) \quad \text{und} \quad g(x+h) = g(x)+J_g(x)h+o(\|h\|)\,.$$

Damit ist auch

$$\begin{aligned}(af + bg)(x + h) &= af(x + h) + bg(x + h)\\ &= af(x) + bg(x) + aJ_f(x)h + bJ_g(x)h + o(\|h\|)\\ &= (af + bg)(x) + \big(aJ_f(x) + bJ_g(x)\big)h + o(\|h\|)\,,\end{aligned}$$

denn auch jede Linearkombination zweier Funktionen der Form $o(\|h\|)$ ist wieder $o(\|h\|)$. Somit ist $af + bg$ differenzierbar mit Ableitung $J_{af+bg}(x) = aJ_f(x) + bJ_g(x)$ für alle $x \in D$.

■

Auch die LEIBNIZ-Regel gilt, d.h.

Lemma: Sind $f, g\colon D \to \mathbb{R}$ stetig differenzierbare Funktionen auf einer Teilmenge $D \subseteq \mathbb{R}^n$, so ist auch $fg\colon D \to R$ differenzierbar und

$$\nabla(fg)(x) = f\nabla g(x) + g\nabla f(x)\,.$$

Beweis: Wir können entweder wie oben mit der Definition argumentieren, oder aber mit partiellen Ableitungen: Nach der Produktregel für Funktionen einer Veränderlichen ist die Produktfunktion für jede der Variablen x_i partiell differenzierbar und

$$\frac{\partial(fg)}{\partial x_i}(x) = f(x)\frac{\partial g}{\partial x_i}(x) + g(x)\frac{\partial f}{\partial x_i}\,.$$

Wegen der vorausgesetzten stetigen Differenzierbarkeit sind diese Ableitungen auch stetig, also ist fg differenzierbar mit diesen Ableitungen als Komponenten des Gradienten. Das sind aber genau die Komponenten von $f\nabla g(x) + g\nabla f(x)$.

■

Schließlich gilt auch eine Kettenregel:

Lemma: $f\colon D \to \mathbb{R}^m$ sei differenzierbar auf $D \subseteq \mathbb{R}^n$ und $g\colon E \to \mathbb{R}^p$ sei differenzierbar auf $E \subseteq \mathbb{R}^m$, wobei $f(D) \subseteq E$ sei. Dann ist auch $g \circ f$ differenzierbar, und $J_{g\circ f}(x) = J_g\big(f(x)\big) J_f(x)$ für alle $x \in D$.

Beweis: Wegen der Differenzierbarkeit von f und g ist für alle $x \in D, y \in E$ und alle $h \in \mathbb{R}^n$, $k \in \mathbb{R}^m$ mit $x + h \in D$ und $y + k \in E$

$$f(x+h) = f(x)+J_f(x)h+o(\|h\|) \quad \text{und} \quad g(y+k) = g(y)+J_g(y)k+o\big(\|k\|\big)\,.$$

Ist $f(x + h) \in E$, gilt daher auch

$$\begin{aligned} g\big(f(x + h)\big) &= g\big(f(x) + J_f(x)h + o(\|h\|)\big) \\ &= g\big(f(x)\big) + J_g\big(f(x)\big)\big(J_f(x)h + o(\|h\|)\big) + \big(J_f(x)h + o(\|h\|)\big) \\ &= g\big(f(x)\big) + J_g\big(f(x)\big) J_f(x)h + o(\|h\|)\,, \end{aligned}$$

denn bei Multiplikation mit der (bezüglich h) konstanten Matrix $J_f(x)$ bleibt eine Funktion $o(\|h\|)$ von dieser Form. ∎

Für weitere Ableitungsregeln sei auf die Übungen verwiesen.

Natürlich gibt es auch im Mehrdimensionalen nicht nur eine Ableitung, sondern wir können Funktionen, entsprechende Differenzierbarkeit vorausgesetzt, auch mehrfach ableiten. Was genau wir unter den höheren Ableitungen einer Funktion mehrerer Veränderlicher verstehen wollen, soll aber erst weiter hinten definiert werden, da wir wegen der bereits zu Beginn dieses Semesters in der Mikroökonomie wichtigen Anwendungen auf Extremwertprobleme mit Nebenbedingungen zunächst solche Probleme betrachten wollen – soweit dies mit der ersten Ableitung allein möglich ist. Auch der nächste Abschnitt ist für uns vor allem wichtig für den Umgang mit Nebenbedingungen.

d) Der Satz über implizite Funktionen

Der Zusammenhang zwischen zwei Größen x und y ist nicht immer explizit in der Form $y = f(x)$ gegeben; gelegentlich hat man auch nur einen impliziten Zusammenhang $F(x, y) = 0$; entsprechend auch für mehr als zwei Variablen. In diesem Abschnitt soll untersucht werden, wann eine Gleichung der Form $F(x) = 0$ nach einer der Variablen x_i aufgelöst werden kann.

In einfachen Fällen ist dies trivial möglich, beispielsweise läßt sich

$$F(x, y, z) = ax + by + cz = 0$$

für $c \neq 0$ durch

$$z = \frac{-ax - by}{c}$$

nach z auflösen. In etwas komplizierteren Fällen, wie etwa bei

$$F(x, y) = x^2 + y^2 - 1 = 0\,,$$

kann man für die Punkte, die nicht auf der x-Achse $y = 0$ liegen, zumindest lokal eindeutig explizit auflösen durch

$$y = \pm\sqrt{1 - x^2}\,,$$

wobei das Vorzeichen gleich dem von y im betrachteten Intervall ist.

Im allgemeinen gibt es jedoch keine Möglichkeit für eine explizite Auflösung mit den „üblichen“ mathematischen Funktionen, d.h. man kann höchstens dann auflösen, wenn man neue Funktionen einführt.

Wie das Beispiel der Kreislinie zeigt, ist auch das nicht immer möglich: Für die beiden Punkte auf der x-Achse gibt es offensichtlich keine *eindeutige* Auflösung, da sowohl die positive wie auch die negative Wurzel Teilauflösungen sind.

Diese Existenz mehrerer Teilauflösungen hängt mit dem Verschwinden der partiellen Ableitung nach y zusammen: Falls diese partielle Ableitung ungleich Null ist, gibt sie an, wie sich F verändert, wenn man y ändert, und sie gibt damit zumindest in erster Näherung auch an, wie man y verändern muß, um bei einer Änderung von x die Bedingung $F(x, y) = 0$ zu erhalten. Falls sie aber verschwindet, fehlt diese Information.

Der Satz über implizite Funktionen besagt, daß das Nichtverschwinden dieser partiellen Ableitung bereits ausreicht um die Existenz einer eindeutigen Auflösung zu zeigen.

Um den Beweis wenigstens einigermaßen überschaubar zu halten, möchte ich mich zunächst auf Funktionen zweier Veränderlicher beschränken:

Satz: $D \subseteq \mathbb{R}^2$ sei offen und $F\colon D \to \mathbb{R}$ sei stetig differenzierbar. Dann gibt es für jeden Punkt $(x_0, y_0) \in D$ mit $F(x_0, y_0) = 0$ und $F_y(x_0, y_0) \neq 0$ Intervallumgebungen I von x_0 und K von y_0 sowie eine eindeutig bestimmte Funktion $f\colon I \to K$, so daß für alle $x \in I$ gilt:

$$F\big(x, f(x)\big) = 0\,.$$

Die Funktion f ist stetig und differenzierbar; ihre Ableitung ist

$$f'(x) = -\frac{F_x(x,y)}{F_y(x,y)} \quad \text{mit} \quad y = f(x)\,.$$

Beweis: Wir beginnen mit einer Reduktion zwecks Vereinfachung der Schreibarbeit: Offensichtlich genügt es, wenn wir den Fall $x_0 = y_0 = 0$ zu betrachten. Gilt nämlich der Satz für die Funktion

$$G(x,y) = F(x + x_0, y + y_0)$$

im Punkt $(0,0)$, so folgt er sofort auch für F im Punkt (x_0, y_0). Außerdem können wir o.B.d.A. annehmen, daß $F_y(0,0)$ positiv ist, denn nach Voraussetzung ist dieser Wert ungleich Null, und falls er negativ sein sollte, ersetzen wir einfach F durch $-F$.

Nach Voraussetzung sind die partiellen Ableitungen von F stetig; daher ist F_y nicht nur im Nullpunkt positiv, sondern auch noch in einer gewissen Umgebung davon. In dieser Umgebung wählen wir ein Rechteck

$$\big\{(x,y) \in \mathbb{R}^2 \;\big|\; -\alpha \leq x \leq \alpha \quad \text{und} \quad -\beta \leq y \leq \beta\big\}\,.$$

Für Punkte aus diesem Rechteck ist $F_y(x,y) > 0$; daher wächst die Funktion $y \mapsto F(x_0, y)$ für jedes $x_0 \in [-\alpha, \alpha]$ streng monoton; wegen $F(0,0) = 0$ ist insbesondere $F(0,-\beta) < 0$ und $F(0,\beta) > 0$. Aufgrund der Stetigkeit von F ist damit für x_1 aus einer gewissen Umgebung der Null auch $F(x_1,-\beta) < 0$ und $F(x_1,\beta) > 0$. Indem wir nötigenfalls α noch etwas verkleinern, können wir annehmen, daß dies sogar für alle $x_1 \in [-\alpha, \alpha]$ gilt.

Damit gibt es nach dem Zwischenwertsatz für jedes $x_1 \in [-\alpha, \alpha]$ ein y_1, so daß $F(x_1, y_1)$ verschwindet; wegen der strengen Monotonie der Funktion $y \mapsto F(x_1, y)$ ist dieser Wert y_1 eindeutig bestimmt. Wir setzen daher $I = (-\alpha, \alpha)$, $K = (-\beta, \beta)$ und

$$f\colon \begin{cases} I & \to K \\ x_1 & \mapsto y_1 \end{cases}\,.$$

Damit ist f als Funktion festgelegt, und nach Konstruktion ist

$$F\big(x, f(x)\big) = 0 \quad \text{für alle } x \in I\,.$$

Wir müssen uns noch überlegen, daß die so konstruierte Funktion f stetig und differenzierbar ist.

Die Stetigkeit können wir etwa dadurch nachweisen, daß wir für jede gegen ein $x \in I$ konvergierende Folge (x_n) aus I zeigen, daß

$$f\big(\lim_{n\to\infty} x_n\big) = \lim_{n\to\infty} f(x_n)$$

ist.

Da x in I liegt, wissen wir, daß es dazu ein eindeutig bestimmtes $y \in K$ gibt, so daß $F(x, y) = 0$ ist, nämlich $y = f(x)$. Für jedes $\varepsilon > 0$ gibt es daher ein $\delta > 0$, so daß $|F(x', y')| < \varepsilon$ ist, falls der Abstand zwischen (x', y') und (x, y) kleiner ist als δ. Zu diesem δ wiederum gibt es ein $N \in \mathbb{N}$, so daß $|x - x_n| < \delta$ für alle $n > N$, so daß für alle solchen n gilt: $|F(x_n, y)| < \varepsilon$. Läßt man hier ε gegen Null gehen, folgt, daß

$$F(x, y) = F(\lim_{n\to\infty} x_n, y) = 0$$

ist, d.h. $F(x, y) = 0$ und damit $f(x) = y$, wie gewünscht.

Somit ist f stetig; als nächstes müssen wir noch die Differenzierbarkeit zeigen. Für ein $x \in I$ und ein hinreichend kleines $h \in \mathbb{R}$, für das auch noch $x + h$ in I liegt, ist nach Definition von f

$$F\big(x + h, f(x + h)\big) = 0\,.$$

Andererseits können wir diesen Funktionswert auch nach dem Mittelwertsatz berechnen: Mit $k = f(x + h) - f(x)$ ist

$$F\big(x + h, f(x + h)\big) = F\big(x + h, f(x) + k\big) = F\left(\big(x, f(x)\big) + \binom{h}{k}\right),$$

und setzen wir

$$\varphi(t) \underset{\text{def}}{=} F\left(\big(x, f(x)\big) + t\binom{h}{k}\right),$$

so ist nach dem Mittelwertsatz der Differentialrechnung

$$\varphi(1) = \varphi(0) + \dot{\varphi}(\tau)$$

für ein τ zwischen Null und eins. Mit $\xi = x + \tau h$ und $\eta = f(x) + \tau k$ ist daher

$$F\big(x + h, f(x) + k\big) = F\big(x, f(x)\big) + hF_x(\xi, \eta) + kF_y(\xi, \eta)\,.$$

Da $F\big(x+h, f(x)+k\big)$ und $F\big(x, f(x)\big)$ beide verschwinden, folgt

$$\frac{f(x+h)-f(x)}{h} = \frac{k}{h} = -\frac{F_x(\xi,\eta)}{F_y(\xi,\eta)}\,.$$

Für $h \to 0$ geht die rechte Seite wegen der Stetigkeit der partiellen Ableitungen gegen $-F_x(x,y)/F_y(x,y)$, insbesondere existiert also der Grenzwert. (Man beachte, daß hier nochmals die Voraussetzung $F_y \neq 0$ benötigt wird). Damit existiert auch der Grenzwert des linksstehenden Differenzenquotienten für $h \to 0$, d.h. f ist differenzierbar und hat die behauptete Ableitung.

∎

Nachdem wir wissen, daß die Ableitung von f existiert, ist ihre Berechnung, unabhängig vom gerade bewiesenen Satz, eine einfache Übungsaufgabe: Die Funktion $F\big(x, f(x)\big)$ ist gleich der Nullfunktion, und damit verschwindet natürlich auch ihre Ableitung. Andererseits ist diese Ableitung nach der Kettenregel gleich

$$F_x\big(x, f(x)\big) + F_y\big(x, f(x)\big) \cdot f'(x)\,,$$

also folgt

$$f'(x) = -\frac{F_x\big(x, f(x)\big)}{F_y\big(x, f(x)\big)}\,.$$

wie Funktionen einer Veränderlichen können auch Funktionen mehrerer Veränderlicher implizit definiert sein; die entsprechende Verallgemeinerung des Satzes über implizite Funktionen folgt fast vollständig aus der koordinatenweisen Anwendung des obigen Satzes, lediglich für die Stetigkeit der Funktion muß man das entsprechende Argument aus dem gerade beendeten Beweis noch einmal anwenden. Die Aussage ist

Satz: $D \subseteq \mathbb{R}^n$ sei eine offene Menge, $f\colon D \to \mathbb{R}$ eine stetig differenzierbare Funktion auf D, und $a = (a_1, \ldots, a_n) \in D$ sei ein Punkt mit $F(a) = 0$. Falls $F_{x_n}(a) \neq 0$ ist, gibt es eine offene Umgebung U von $(a_1, \ldots, a_{n-1})$ in $\mathbb{R}^{n-1}$ und eine Funktion $f \in \mathcal{C}^1(U, \mathbb{R})$, so daß für alle Punkte $y = (y_1, \ldots, y_{n-1}) \in U$ gilt: $F\big(y, f(y)\big) = 0$. Für alle $i \leq n-1$ ist

$$f_{x_i}(y) = -\frac{F_{x_i}\big(y, f(y)\big)}{F_{x_n}\big(y, f(y)\big)}\,.$$

∎

e) Ableitungen und Extrema

Im Eindimensionalen ist das Verschwinden der Ableitung eine notwendige Bedingung für einen Extremwert. Dies gilt genau so auch im Mehrdimensionalen:

Für eine differenzierbare Funktion $f\colon D \to \mathbb{R}$ auf einer offenen Teilmenge $D \subset \mathbb{R}^n$ ist im Punkt $x_0 \in D$

$$f(x_0 + h) = f(x_0) + \langle \nabla f(x_0), h \rangle + o(\|h\|)\,.$$

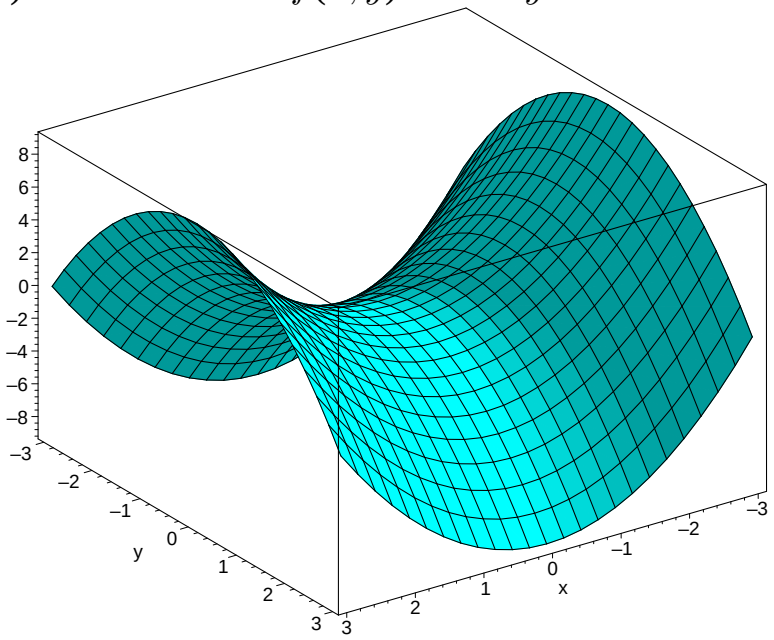
Daher muß für jeden Extremwert grad $f(x_0)$ gleich dem Nullvektor sein, denn setzt man für h ein kleines Vielfaches $t \cdot \text{grad}\, f(x_0)$ des Gradienten ein, wäre sonst

$$f(x_0 + h) = f(x) + t\,\langle \nabla f(x_0), \nabla f(x_0) \rangle + o(\|h\|)$$

für kleine positive t größer als $f(x_0)$ und für kleine negative t kleiner.

Die Frage, welche Nullstellen des Gradienten wirklich Extremwerten entsprechen, ist schwieriger; in der Praxis wird es oft am einfachsten sein, sich die Umgebung des betreffenden Punktes mit irgendwelchen *ad hoc*-Methoden genauer anzusehen und dann zu entscheiden.

Klassisches Beispiel eines Punkts, in dem der Gradient verschwindet, ohne daß ein Extremwert vorliegt, ist der in der folgenden Abbildung gezeigte Sattelpunkt, hier dargestellt als Funktionswert über dem Punkt $(0, 0)$ für die Funktion $f(x, y) = x^2 - y^2$.



Graph der Funktion $f(x,y) = x^2 - y^2$

Wie wir bald sehen werden, kann man auch hier mit den (noch zu definierenden) höheren Ableitungen auch hinreichende Bedingungen finden; diese sind allerdings schon für Funktionen von zwei oder drei Veränderlichen deutlich aufwendiger als im Falle einer Veränderlichen.

f) Extremwerte unter Nebenbedingungen

Bei einem realen physikalischen, technischen wirtschaftlichen oder sozialen Prozeß können sich die Variablen selten frei im gesamten $\mathbb{R}^n$ bewegen: Sinnvoll und realistisch ist meist nur eine beschränkte Teilmenge. Im Gegensatz zur Dimension eins, wo diese Teilmenge praktisch immer ein Intervall ist, kann sie im Mehrdimensionalen durch Randbedingungen aller Art charakterisiert sein und damit auch beliebig kompliziert aussehen.

Maxima und Minima auf solchen Teilmengen sind im allgemeinen keine lokalen Maxima oder Minima der betrachteten Funktion: Wenn man die jeweilige Fläche verläßt, läßt sich der Funktionswert selbst für einen solchen Extremwert meist noch – je nach Richtung – sowohl vergrößern als auch verkleinern. Dementsprechend können die Methoden, die wir im vorigen Abschnitt diskutiert haben, solche Extremwerte üblicherweise nicht finden. Wir brauchen daher neue Werkzeuge, und die soll dieser Paragraphen bereitstellen.

Die Ausgangslage ist typischerweise die folgende: Gegeben ist eine Funktion $f\colon \mathbb{R}^n \to \mathbb{R}$, möglicherweise auch nur auf einer Teilmenge $D \subset \mathbb{R}^n$ definiert, deren Extremwerte nicht auf $\mathbb{R}^n$ oder D gesucht werden, sondern nur auf einer Teilmenge, die beispielsweise durch das Verschwinden einer weiteren Funktion $g\colon \mathbb{R}^n \to \mathbb{R}$ gegeben ist. Falls wir uns für Extremwerte auf einer Kugel vom Radius r um den Nullpunkt interessieren, wäre dies etwa die Funktion

$$g\colon \begin{cases} \mathbb{R}^3 & \to \mathbb{R} \\ (x,y,z) & \mapsto x^2 + y^2 + z^2 - r^2 \end{cases}.$$

Eine mögliche Strategie zur Lösung solcher Probleme besteht darin, die Gleichung $g = 0$ nach einer der Variablen aufzulösen, diese dann in f einzusetzen und sodann eine gewöhnliche Extremwertaufgabe zu lösen.

Diese Auflösung ist *explizit* nur in sehr einfachen Fällen möglich, aber selbst wenn wir nur wissen, daß eine solche Auflösung *existiert,* können wir doch damit argumentieren und Kriterien ableiten.

Unter Maxima und Minima sollen hier *lokale* Extrema verstanden werden, so daß wir die üblichen Kriterien anwenden können:

Definition: Wir sagen, die Funktion $f\colon D \to \mathbb{R}$ auf einer Teilmenge $D \subseteq \mathbb{R}^n$ habe im Punkt $a \in D$ ein lokales $\left\{\begin{matrix}\text{Maximum}\\ \text{Minimum}\end{matrix}\right\}$ unter der Nebenbedingung $g = 0$, wobei $g\colon D \to \mathbb{R}$ eine weitere Funktion ist, wenn $g(a) = 0$ ist und es eine Umgebung U von a gibt, so daß für alle $x \in U$ gilt: Ist $g(x) = 0$, so ist $f(x) \left\{\begin{matrix}\leq\\ \geq\end{matrix}\right\} f(a)$.

Als Einstiegsbeispiel betrachten wir eine beliebte Schulbuchaufgabe zur Minimumsbestimmung: Eine Konservendose soll bei einem vorgegebenen Volumen von 100 cm^3 möglichst wenig Blech benötigen, d.h. ihre Oberfläche soll minimal sein.

Die Oberfläche eines Zylinders der Höhe h mit einer Grundfläche vom Radius r ist

$$f(r,h) = 2\pi r^2 + 2\pi r \cdot h\,;$$

wegen der Nebenbedingung für das Volumen $V = \pi r^2 h$ muß gelten

$$g(r,h) = \pi r^2 h - 100 = 0\,.$$

Hier läßt sich natürlich die Nebenbedingung sofort nach h auflösen:

$$h = \frac{100}{\pi r^2}\,,$$

und wir müssen nur noch die Funktion

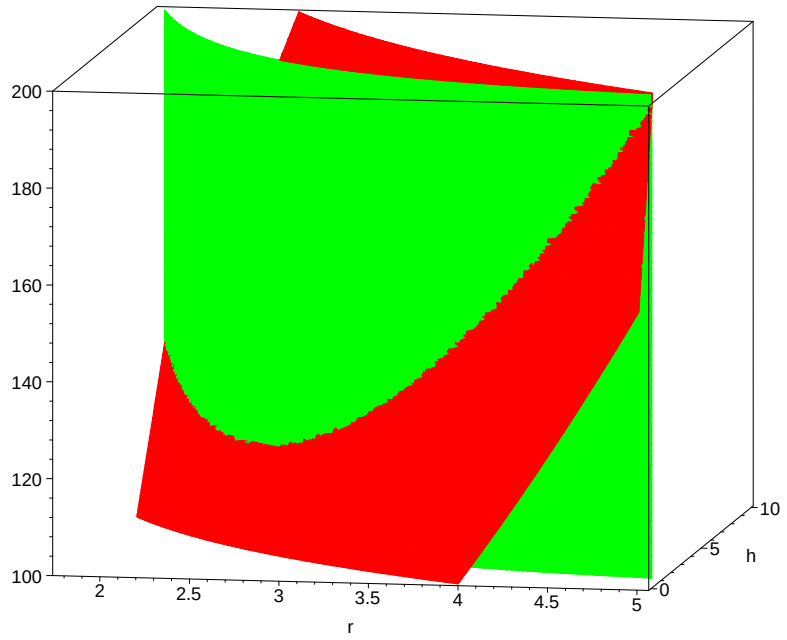
$$F(r) = f\left(r, \frac{100}{\pi r^2}\right) = 2\pi r^2 + \frac{200}{r}$$

minimieren. Für diese ist

$$F'(r) = 4\pi r - \frac{200}{r^2}\,,$$

und dies verschwindet genau dann, wenn gilt

$$4\pi r^3 = 200 \quad \text{oder} \quad r = \sqrt[3]{\frac{50}{\pi}}\,.$$



Oberfläche einer Konservendose mit festem Volumen

In diesem einfachen Fall konnten wir die Optimierung also zurückführen auf gewöhnliche Extremwertaufgaben einer Veränderlichen, indem wir die Nebenbedingung nach einer der Variablen auflösten und diese dann in f einsetzten. Im allgemeinen wird dies aber nicht möglich sein, so daß wir andere Methoden brauchen.

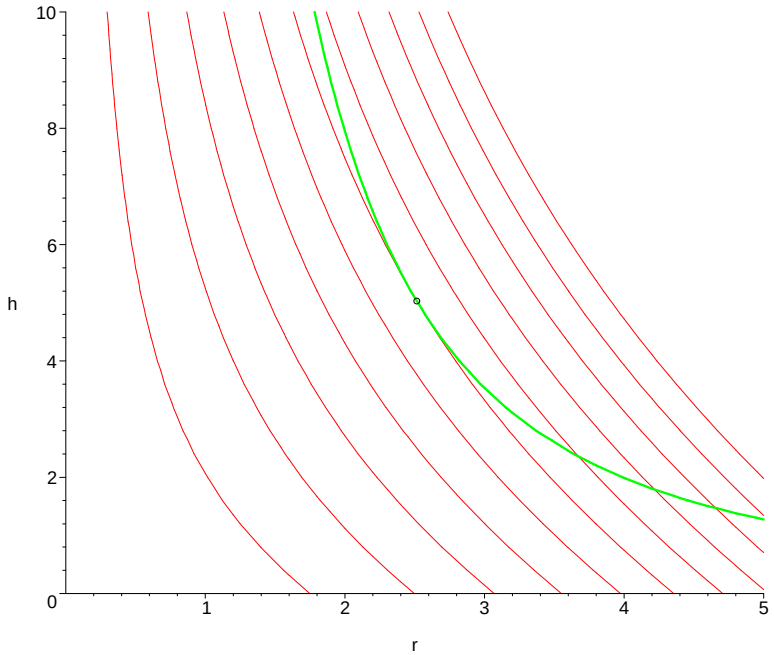
Unser bisherige Theorie für lokale Extrema ist in dieser Situation nicht anwendbar, denn die lokalen Extrema von f werden nur in den seltensten Fällen die Nebenbedingung $g = 0$ erfüllen.

In der obigen Abbildung ist die Nebenbedingung als grüne Fläche dargestellt und der Graph von f als rote; wie man sieht, läßt sich der Wert von f problemlos verkleinern, wenn man nur die Fläche $g = 0$ verläßt, und in der Tat ist auch ohne jede Mathematik sofort klar, daß man mit weniger Blech auskommt, wenn man die Konservendose einfach schmaler oder kürzer macht.

Die Grundidee für ein alternatives Verfahren wird klar bei der Betrachtung der Niveaulinien: Die Niveaulinie für $g = 0$ ist grün eingezeichnet, verschiedene Niveaulinien von f als rote Kurven.

Wie man sieht, schneiden einige dieser Niveaulinien die gestrichelte Kurve überhaupt nicht: Wenn man zu wenig Blech hat, kann man keine

Dose mit 100 cm^3 Inhalt zusammenlöten. Wenn es dagegen genug Blech gibt, gibt es gleich zwei Schnittpunkte: Die Dose kann entweder eher höher oder eher breiter gemacht werden. In einem solchen Fall kann man die Niveaulinie durch eine zu einem etwas niedrigeren Niveau ersetzen, die im allgemeinen auch wieder Schnittpunkte haben wird, so daß das Niveau noch nicht minimal sein kann. Erst wenn man im Minimum ist, fallen die beiden Schnittpunkte zusammen; wenn man nun das Niveau noch weiter erniedrigt, gibt es keine Schnittpunkte mehr.



Niveaulinien für Oberfläche und Volumen

Da somit im Minimum zwei Schnittpunkte zusammenfallen, berühren sich dort die Niveaulinien von f und von g, d.h. sie haben eine gemeinsame Tangente. Da der Gradient, wie wir wissen, senkrecht auf der Tangenten der Niveaulinien steht (die Richtungsableitung entlang einer Niveaulinie ist schließlich Null), sind somit die Gradienten von f und g im Minimum zueinander parallel, d.h. der eine ist ein Vielfaches des anderen.

Dies gilt nicht nur im vorliegenden Beispiel, sondern allgemein:

Satz: $D \subseteq \mathbb{R}^n$ sei eine offene Menge und $f, g \in \mathcal{C}^1(D, \mathbb{R})$ seien stetig differenzierbare Funktionen auf D. Falls f im Punkt $a \in D$ ein Extremum hat unter der Nebenbedingung $g(x) = 0$, so sind grad $f(a)$ und grad $g(a)$ linear abhängig.

Beweis: Die Grundidee ist einfach: Auch wenn wir die Nebenbedingung nicht *explizit* nach einer der Variablen auflösen können, sagt uns der Satz über implizite Funktionen in vielen Fällen dennoch, daß zumindest lokal eine Auflösung existiert. Diese Auflösung kennen wir zwar nicht, aber wir können mit ihr argumentieren und, zumindest formal, auch rechnen.

Falls grad $g(a)$ der Nullvektor ist, gibt es nichts mehr zu beweisen, denn jede Menge, die den Nullvektor enthält, ist linear abhängig.

Wir können daher annehmen, daß grad $g(a)$ mindestens eine von Null verschiedene Komponente hat, und durch Umnummerieren der Koordinaten können wir o.B.d.A. annehmen, daß dies die n-te Komponente ist, d.h. $g_{x_n}(a) \neq 0$.

Dann gibt es nach dem Satz über implizite Funktionen eine Umgebung U von $(a_1, \dots, a_{n-1})$ sowie eine Funktion $h\colon U \to \mathbb{R}$ mit $h(a_1, \dots, a_{n-1}) = a_n$, so daß

$$g\big(x_1, \dots, x_{n-1}, h(x_1, \dots, x_{n-1})\big) = 0 \quad \text{für alle } (x_1, \dots, x_{n-1}) \in U\,.$$

Nachdem f in a ein lokales Extremum unter der Nebenbedingung $g = 0$ hat, nimmt die Funktion

$$F(x_1, \dots, x_{n-1}) \underset{\text{def}}{=} f\big(x_1, \dots, x_{n-1}, h(x_1, \dots, x_{n-1})\big)$$

in $(a_1, \dots, a_{n-1})$ ein lokales Extremum im üblichen Sinne an, d.h. der Gradient von F verschwindet dort.

Nach der Kettenregel ist für $i = 1, \dots, n-1$

$$F_{x_i}(a_1, \dots, a_{n-1}) = f_{x_i}(a) + f_{x_n}(a) \cdot h_{x_i}(a_1, \dots, a_{n-1})\,,$$

und nach dem Satz über implizite Funktionen ist $h_{x_i} = -g_{x_i}/g_{x_n}$, d.h.

$$F_{x_i}(a_1, \dots, a_{n-1}) = f_{x_i}(a) - f_{x_n}(a)\frac{g_{x_i}(a)}{g_{x_n}(a)}\,.$$

Da die linke Seite verschwindet, gilt dasselbe auch für die rechte. Die rechte Seite ist im Gegensatz zur linken auch für $i = n$ definiert und verschwindet aus trivialen Gründen; also ist für alle i

$$f_{x_i}(a) - \frac{f_{x_n}(a)}{g_{x_n}(a)} g_{x_i}(a) = 0$$

oder, anders ausgedrückt, grad $f(a) - \dfrac{f_{x_n}(a)}{g_{x_n}(a)}$ grad $g(a) = 0$. Damit sind die beiden Gradienten in der Tat linear abhängig. ∎

Falls der Gradient von g im Punkt a nicht verschwindet, gibt es somit eine Zahl $\lambda \in \mathbb{R}$, so daß grad $f(a) - \lambda$ grad $g(a) = 0$ ist, nämlich

$$\lambda = \frac{f_{x_n}(a)}{g_{x_n}(a)} .$$

Diese Zahl bezeichnet man als LAGRANGEschen Multiplikator; mit seiner inhaltlichen Interpretation werden wir uns in Kürze beschäftigen.

JOSEPH-LOUIS LAGRANGE (1736–1813) wurde als GIUSEPPE LODOVICO LAGRANGIA in Turin geboren und studierte dort zunächst Latein. Erst eine alte Arbeit von HALLEY über algebraische Methoden in der Optik weckte sein Interesse an der Mathematik, woraus ein ausgedehnter Briefwechsel mit EULER entstand. In einem Brief vom 12. August 1755 berichtete er diesem unter anderem über seine Methode zur Berechnung von Maxima und Minima; 1756 wurde er, auf EULERs Vorschlag, Mitglied der Berliner Akademie; zehn Jahre später zog er nach Berlin und wurde dort EULERs Nachfolger als mathematischer Direktor der Akademie. 1787 wechselte er an die Pariser Académie des Sciences, wo er bis zu seinem Tod blieb und unter anderem an der Einführung des metrischen Systems beteiligt war. Seine Arbeiten umspannen weite Teile der Analysis, Algebra und Geometrie.

Zur praktischen Bestimmung von Extremwerten unter Nebenbedingungen geht man wie folgt vor: Über die Punkte, in denen der Gradient von g verschwindet, macht obiger Satz keine verwertbare Aussage; diese Punkte müssen also vorab berechnet und untersucht werden.

Danach müssen die Punkte gefunden werden, in denen es ein $\lambda \in \mathbb{R}$ gibt, so daß

$$\begin{aligned} f_{x_1}(x) - \lambda g_{x_1}(x) &= 0 \\ &\vdots \\ f_{x_n}(x) - \lambda g_{x_n}(x) &= 0 \\ g(x) &= 0 \end{aligned}$$

ist. Mit der LAGRANGE-Funktion

$$\mathcal{L}(x_1, \ldots, x_n, \lambda) = f(x) - \lambda g(x)$$

läßt sich dies auch kurz schreiben als $\nabla\mathcal{L}(x_1, \ldots, x_n, \lambda) = 0$, denn die partiellen Ableitungen von $\mathcal{L}$ sind abgesehen vom Vorzeichen der Ableitung nach λ gerade die linken Seiten diesen Gleichungssystems.

Dieses ist ein System von $n + 1$ Gleichungen für $n + 1$ Unbekannte, allerdings ist es nur selten linear und damit oft nicht mit den uns bekannten Methoden lösbar. Manchmal kann man das Gleichungssystem durch geeignete Umformungen und Fallunterscheidungen vollständig lösen, in anderen Fällen helfen nur die aus der Numerik bekannten Näherungsverfahren wie etwa die Methode von NEWTON-RAPHSON.

Falls alle Gleichungen Polynomgleichungen sind (oder durch Einführung geeigneter zusätzlicher Variablen auf Polynomgleichungen zurückgeführt werden können), kann man im Falle einer endlichen Lösungsmenge diese auch exakt bestimmen: Genau wie der GAUSS-Algorithmus zur Lösung eines linearen Gleichungssystems dieses auf eine Treppengestalt bringt, aus der man die Lösungen einfach ermitteln kann, gibt es in der Computeralgebra einen Algorithmus, der dasselbe für beliebige Systeme von Polynomgleichungen versucht; die Gleichungen, die dieser Algorithmus liefert, bezeichnet man als GRÖBNER-Basis oder Standardbasis. Zum Verständnis dieses Algorithmus, den man als eine Art Synthese aus EUKLIDischen Algorithmus und GAUSS-Algorithmus ansehen kann, sind Kenntnisse der kommutativen Algebra erforderlich, für die die Zeit in dieser Vorlesung nicht ausreicht; bei einigen Implementierungen werden zusätzlich auch noch Algorithmen aus der Informatik eingesetzt, die typischerweise nicht in Grundvorlesungen behandelt werden. Deshalb sei hier nur darauf hingewiesen, daß die gängigen universellen Computeralgebrasysteme wie Maple, Mathematica, MuPad allesamt entsprechende Routinen enthalten, mit denen man auch dann experimentieren kann, wenn man die dahinter stehende Theorie nicht versteht.

Als Beispiel, wie gelegentlich auch ein nichtlineares Gleichungssystem elementar gelöst werden kann, betrachten wir eine Anwendung aus den Wirtschaftswissenschaften: Die Gesamtproduktion eines Unternehmens oder eines Staats in Abhängigkeit von n eingesetzten Ressourcen $x_1, \ldots, x_n$ wird oft modelliert durch eine sogenannte COBB-DOUGLAS-Funktion der Form

$$P(x_1, \ldots, x_n) = \alpha x_1^{e_1} \ldots x_n^{e_n} ,$$

benannt nach den beiden Wissenschaftlern, die dieses Modell 1928 für die amerikanische Gesamtproduktion in Abhängigkeit von Kapital und Arbeit in den Jahren 1899 bis 1922 entwickelten. (Sie fan-

den $P \approx 1{,}01A^{3/4}K^{1/4}$ mit A = Anzahl der Beschäftigten und K = Kapitaleinsatz.)

Betrachten wir stattdessen die Produktion eines Wirtschaftsguts aus zwei Ressourcen x, y gemäß der Funktion

$$f(x,y) = P(x,y) = x^{1/2}y^{1/4}\,.$$

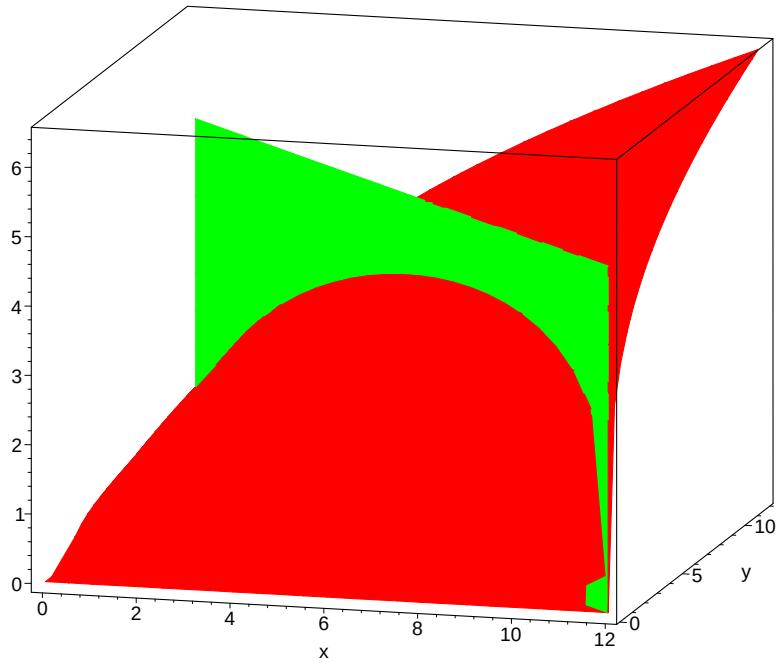
Falls wir der Einfachheit halber annehmen, daß die Kosten pro Einheit für x und y gleich sind und die Gesamtkosten höchstens gleich zwölf sein dürfen, müssen wir f maximieren unter der Nebenbedingung

$$x + y \leq 12\,.$$

Nun ist aber f eine monoton wachsende Funktion sowohl von x als auch von y, d.h. die maximale Produktion wird sicherlich erreicht in einem Punkt, für den $x + y = 12$ ist, denn für jeden anderen Punkt (x,y) mit $x + y < 12$ ist $f(x,y) < f(x, 12-x)$. Daher können wir die Nebenbedingung in der gewohnten Form

$$g(x,y) = x + y - 12 = 0$$

schreiben.



Maximierung einer Produktionsfunktion bei festem Kapitaleinsatz

Ableitung beider Funktionen zeigt, daß

$$\operatorname{grad} g = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \text{und} \quad \operatorname{grad} f = \begin{pmatrix} y^{1/4}/2x^{1/2} \\ x^{1/2}/4y^{3/4} \end{pmatrix}$$

ist; das zu lösende Gleichungssystem wird also zu

$$\begin{aligned} \frac{y^{1/4}}{2x^{1/2}} - \lambda &= 0 \\ \frac{x^{1/2}}{4y^{3/4}} - \lambda &= 0 \\ x + y - 12 &= 0 \end{aligned} .$$

(Die Nenner brauchen uns nicht zu stören, denn da $f(0, y) = f(x, 0) = 0$ ist, kommen Lösungen mit $x = 0$ oder $y = 0$ für das Maximum ohnehin nicht in Betracht; wir können sie also getrost ausschließen.)

Als Ansatz zu einer möglichen Lösung können wir ausnutzen, daß λ in den beiden ersten Gleichungen isoliert steht; wenn wir danach auflösen und gleichsetzen, erhalten wir die Gleichung

$$\frac{y^{1/4}}{2x^{1/2}} = \frac{x^{1/2}}{4y^{3/4}} .$$

Multiplikation mit dem Hauptnenner macht daraus

$$4y^{1/4}y^{3/4} = 2x^{1/2}x^{1/2} \quad \text{oder} \quad 2y = x .$$

Einsetzen in die dritte Gleichung ergibt $3y = 12$, also ist

$$y = 4 \quad \text{und} \quad x = 8 ;$$

der Maximalwert von f ist

$$f(8, 4) = 8^{1/2} \cdot 4^{1/4} = 2\sqrt{2} \cdot \sqrt{2} = 4 .$$

Auch den LAGRANGEschen Multiplikator λ können wir noch ausrechnen:

$$\lambda = \frac{y^{1/4}}{2x^{1/2}} = \frac{4^{1/4}}{2 \cdot 8^{1/2}} = \frac{\sqrt{2}}{2 \cdot 2\sqrt{2}} = \frac{1}{4} .$$

Die Berechnung von λ war für die Bestimmung des Optimums eigentlich überflüssig; λ ist nur eine Hilfsgröße zur Berechnung des Extremums.

Wir wollen uns als nächstes überlegen, daß wir λ auch inhaltlich interpretieren können: Dazu betrachten wir eine Nebenbedingung

$$g(x_1, \ldots, x_n) = c$$

mit *variabler* rechter Seite c und ein Extremum der Funktion

$$f(x_1, \ldots, x_n)\,.$$

Dieses Extremum wird natürlich von c abhängen; wir schreiben es in der Form

$$\big(x_1(c), \ldots, x_n(c)\big)$$

und nehmen an, daß die Funktionen $x_i(c)$ stetig differenzierbar seien. (Ein interessierter Leser kann sich anhand des Satzes über implizite Funktionen überlegen, welche Bedingungen f und g erfüllen müssen, damit dies garantiert ist.) Der Optimalwert von f in Abhängigkeit von c ist dann

$$F(c) \underset{\text{def}}{=} f\big(x_1(c), \ldots, x_n(c)\big)\,.$$

Nach der Kettenregel ist

$$F'(c) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \frac{dx_i(c)}{dc}\,.$$

Genauso folgt für $G(c) \underset{\text{def}}{=} g\big(x_1(c), \ldots, x_n(c)\big)$ die Gleichung

$$G'(c) = \sum_{i=1}^{n} \frac{\partial g}{\partial x_i} \frac{dx_i(c)}{dc}\,.$$

Da $\big(x_1(c), \ldots, x_n(c)\big)$ ein Optimum ist, sind dort die Gradienten von f und $g - c$ proportional mit Proportionalitätsfaktor λ, und da wir bei der Gradientenbildung nur nach den x_i ableiten, von denen die rechte Seite c nicht abhängt, ist der Gradient von $g - c$ gleich dem von g selbst, d.h.

$$\frac{\partial f}{\partial x_i} = \lambda \frac{\partial g}{\partial x_i} \qquad \text{für alle } i\,.$$

Somit ist $F'(c) = \lambda G'(c)$. Nun erfüllt aber der Punkt $\big(x_1(c), \ldots, x_n(c)\big)$ die Nebenbedingung mit rechter Seite c; somit ist $G(c) = c$ und damit $G'(c) \equiv 1$. Damit ist $\lambda = F'(c)$ die Wachstumsrate für das Optimum bei Änderung der rechten Seite der Nebenbedingung.

Im obigen Beispiel steigt also die Maximalmenge $f(x, y)$, die mit Kapitaleinsatz 12 produziert werden kann, für kleines h ungefähr um $h/4$, wenn wir den Kapitaleinsatz auf 12 + h erhöhen. Die Erhöhung des Kapitaleinsatzes lohnt sich, wenn für das fertige Produkt ein Preis pro Einheit erzielt werden kann, der größer ist als vier.

Als letztes wollen wir uns noch überlegen, was passiert, wenn wir nicht nur eine, sondern mehrere Nebenbedingungen erfüllen müssen. Es geht also wieder darum, eine Funktion $f\big(x_1, \ldots, x_n\big)$ zu optimieren, jetzt aber unter den Nebenbedingungen

$$g_1(x_1, \ldots, x_n) \geq 0, \quad \ldots \quad g_r(x_1, \ldots, x_n) \geq 0\,.$$

(Es genügt, Bedingungen mit $\geq$ zu betrachten, denn durch Multiplikation mit minus Eins kann man jede Ungleichung mit $\leq$ in eine mit $\geq$ überführen. Auch Gleichungen $g_i = 0$ kann man zumindest formal durch die beiden Ungleichungen $g_i \geq 0$ und $-g_i \geq 0$ ausdrücken.)

Die wichtigsten Beispiele solcher Optimierungsaufgaben sind die Fälle mit linearen Funktionen f und g_i; hier redet man von *linearen Programmen*. (Das Wort *Programm* in diesem Zusammenhang hat natürlich nichts mit Computerprogrammen zu tun.) Das wichtigste Verfahren zur Lösung solcher Aufgaben, der Simplex-Algorithmus, wird in der Vorlesung *Diskrete Mathematik A* behandelt, so daß wir uns hier auf die *nichtlineare Programmierung* beschränken können.

Man überlegt sich leicht, daß im linearen Fall die Nebenbedingungen ein (endliches oder unendliches) Polyeder im $\mathbb{R}^n$ definieren und eine lineare Funktion, so sie ein endliches Maximum oder Minimum hat, dieses auf dem Rand dieses Polyeders annimmt, und dort sogar in einer Ecke. Man muß daher „nur“ die Ecken dieses Polyeders untersuchen – deren Anzahl allerdings wächst exponentiell mit der Anzahl der Variablen. Trotzdem führt der Simplex-Algorithmus selbst im Fall von Zehntausenden von Variablen in der Regel fast immer sehr schnell ans Ziel; das theoretische Problem der exponentiellen Komplexität im schlimmsten Fall hat also für praktische Anwendungen keine Bedeutung.

Bei nichtlinearen Funktionen ist die Situation komplizierter, denn nun

kann es auch im Innern Extrema geben: Die Funktion

$$f(x,y) = e^{-x^2-y^2} \quad \text{mit der Nebenbedingung} \quad x^2 + y^2 \leq 1$$

etwa nimmt ihr Maximum im Punkt $(0,0)$ an; auf dem Rand des Einheitskreises liegen nur die Minima. Im allgemeinen Fall eines nichtlinearen Programms kann ein Optimum also entweder ganz im Innern liegen oder aber eine beliebige Teilmenge der Nebenbedingungen exakt erfüllen.

Falls wir es mit inneren Punkte zu tun haben, sind diese lokale Maxima oder Minima ohne Nebenbedingungen, und wir haben uns bereits in §1 überlegt, wie man diese bestimmt: In jedem solchen Punkt verschwindet der Gradient der zu optimierenden Funktion.

Im Falle einer einzigen *Gleichung* als Nebenbedingung ist der Gradient von f linear abhängig vom Gradienten der Nebenbedingung; da der Nullvektor von jedem anderen Vektor linear abhängig ist, schließt dies auch den Fall der Optima bei inneren Punkten mit ein. Die naheliegende Verallgemeinerung auf den Fall mehrerer Nebenbedingungen ist der

Satz: Die Funktion $f\colon D \to \mathbb{R}$ auf $D \subseteq \mathbb{R}^n$ habe im Punkt $a \in D$ ein Extremum unter den Nebenbedingungen

$$g_1(a) \geq 0, \quad g_2(a) \geq 0, \quad \ldots, \quad g_r(a) \geq 0\,.$$

Dann sind die $r+1$ Vektoren

$$\nabla f(a), \quad \nabla g_1(a), \quad \nabla g_2(a), \quad \ldots, \quad \nabla g_r(a)$$

linear abhängig.

Der *Beweis* erfordert keine wesentlich neuen Ideen gegenüber dem Fall einer einzigen Nebenbedingung und sei daher nur kurz skizziert: Falls die Gradienten der g_i im Punkt a bereits untereinander linear abhängig sind, gibt es nichts mehr zu beweisen; nehmen wir also an, sie seien linear unabhängig. Dann gibt es (mindestens) r verschiedene Variablen x_{j_1} bis x_{j_r}, so daß

$$\frac{\partial g_i}{\partial x_{j_i}}(a) \neq 0$$

ist. Also kann nach dem Satz über implizite Funktionen jede Nebenbedingung zur Elimination einer anderen Variablen benutzt werden, und

im wesentlichen dieselbe Rechnung wie im Fall einer Nebenbedingung zeigt die Behauptung.

■

Die lineare Abhängigkeit der Vektoren

$$\nabla f(a), \quad \nabla g_1(a), \quad \nabla g_2(a), \quad \ldots, \quad \nabla g_r(a)$$

bezeichnet man als KUHN-TUCKER-Bedingung; sie ist eine offensichtliche Verallgemeinerung der Bedingung von LAGRANGE, ist allerdings deutlich jünger: Sie erschien 1951 in einer gemeinsamen Arbeit von H.W. KUHN und A.W. TUCKER, vier Jahre, nachdem G. DANTZIG den Simplex-Algorithmus entwickelt hatte, und fast zweihundert Jahre, nachdem LAGRANGE seine Multiplikatoren zur Bestimmung von Extrema unter einer Nebenbedingung eingeführt hatte.

Das Problem bei der praktischen Anwendung des Satzes von KUHN und TUCKER besteht darin, daß in einem Optimum manche Nebenbedingungen als Gleichungen, andere als echte Ungleichungen erfüllt sind; man muß also jede der möglichen Kombinationen untersuchen.

Eine mögliche Abhilfe sind sogenannte *barrier*-Methoden: Man läßt die Nebenbedingungen eine Barriere errichten, indem man (bei der Suche nach einem Maximum) Maxima *ohne* Nebenbedingung der Funktion

$$f(x_1, \ldots, x_n) + \sum_{i=1}^{r} \varepsilon_i \log g_i(x_1, \ldots, x_n)$$

sucht, wobei die ε_i positive Konstanten sind. Da die Logarithmen am Rand gegen $-\infty$ gehen, liegen diese Maxima stets im Innern. Falls man nun alle ε_i in geeigneter Weise gegen Null gehen läßt, kann man in manchen Fällen zeigen, daß diese Maxima gegen Maxima der Funktion *mit* Nebenbedingung konvergiert.

Ein Beispiel dafür ist der 1984 gefundene Algorithmus von KARMAKAR für den Fall linearer Funktionen f, g_i. Er ist eine Alternative zum Simplex-Algorithmus, die stets in polynomialer Zeit zu einer Lösung führt, und war der erste mathematische Algorithmus, der patentiert wurde. In der Praxis ist er jedoch bei fast allen Problemen dem Simplex-Algorithmus unterlegen; lediglich bei einigen wenigen Spezialfällen, bei denen bekannt ist, daß der Simplex-Algorithmus schlecht funktioniert, führt KARMAKAR schneller zu einer Lösung.

g) Ausblick: Numerische Methoden

Wie wir gesehen haben, führt die Methode der LAGRANGEschen Multiplikatoren im allgemeinen auf nichtlineare Gleichungssysteme, die nur in einfachen Fällen explizit lösbar sind. In allen anderen Fällen muß man mit numerischen Methoden arbeiten, und da bietet sich an, das Problem von vornherein ohne den Umweg über LAGRANGEsche Multiplikatoren Extrema numerisch zu bearbeiten.

Eine Möglichkeit dazu ist die sogenannte *Gradientenmethode:*

Für eine differenzierbare Funktion f auf $D \subseteq \mathbb{R}^n$ ist

$$f(x+h) = f(x) + \langle \text{grad } f(x), h\rangle + o(\|h\|)\,;$$

wenn wir ein Maximum (oder Minimum) von f ansteuern wollen, liegt es daher nahe, h so zu wählen, daß sich der Funktionswert möglichst stark vergrößert (oder verkleinert).

Nach der CAUCHY-SCHWARZschen Ungleichung ist, wenn wir mit der EUKLIDischen Norm arbeiten,

$$|\langle \text{grad } f(x), h\rangle| \leq \|\text{grad } f(x)\| \cdot \|h\|\ ;$$

wir erhalten also die maximal mögliche Veränderung bei vorgegebener Länge von h genau dann, wenn h parallel zum Gradienten ist.

Damit bietet sich folgende Strategie an: Wir wählen irgendeinen Ausgangspunkt x_0 und berechnen dort den Gradienten $\nabla f(x_0)$. Falls er der Nullvektor ist. haben wir einen Kandidaten für ein Extremum gefunden, den wir mit noch zu entwickelnden Methoden weiter untersuchen müssen.

Andernfalls geben wir uns eine Länge ℓ_0 für den Vektor h vor, die von der Länge des Gradienten abhängen kann oder auch nicht, und setzen wir bei der Suche nach einem Maximum

$$h_0 = \frac{\ell_0}{\|\nabla f(x_0)\|}\nabla f(x_0)\,;$$

bei der Suche nach Minima nehmen wir das Negative davon.

Als nächstes betrachten wir den Punkt

$$x_1 \underset{\text{def}}{=} x_0 + h_0\,,$$

berechnen dort den Gradienten $\nabla f(x_1)$, setzen – so er nicht verschwindet – mit einer geeigneten Länge ℓ_1

$$h_1 = \pm \frac{\ell_1}{\|\nabla f(x_1)\|} \nabla f(x_1)$$

(+ für Maxima, $-$ für Minima) zur Definition des nächsten Punkts

$$x_2 \underset{\text{def}}{=} x_1 + h_1$$

und so weiter. In jedem Schritt erhöhen (oder erniedrigen) wir den Funktionswert soweit, wie es mit der vorgegebenen Länge ℓ_i nur möglich ist, in der Hoffnung, so irgendwann auf ein Maximum (oder Minimum) zu stoßen. Dieses können wir erreichen, wenn wir am Rand des Definitionsbereichs von f angelangt sind, oder aber wenn wir in einem Punkt sind, in dem der Gradient verschwindet: Von dort aus geht es mit diesem Verfahren nicht mehr weiter.

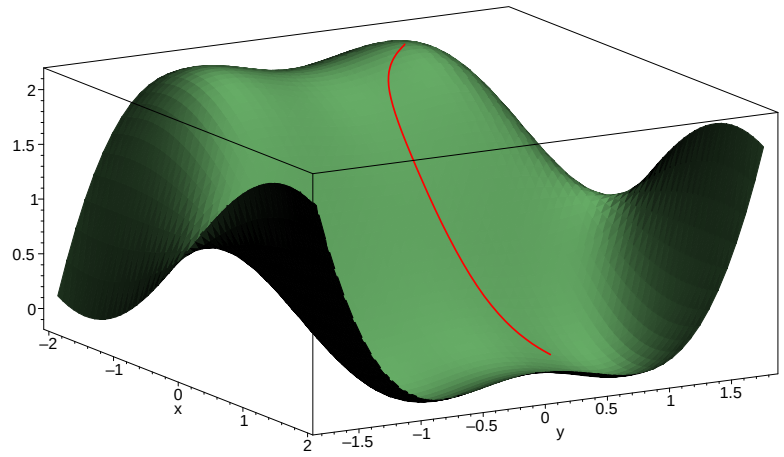
Da wir mit einem numerischen Verfahren nur ein verschwindend geringe Chance haben, exakt in einem Extremum zu enden, zeigt sich hier auch die Notwendigkeit einer intelligenten Wahl der Schrittweiten ℓ_i: Wenn diese zu groß sind, kann es passieren, daß wir endlos um ein Extremum herum oszillieren.

Theoretisch ist auch möglich, daß wir in einem Sattelpunkt landen, aber wenn man sich überlegt, wie die Gradienten in der Umgebung eines Sattelpunktes aussehen, wird schnell klar, daß dies nur sehr selten passiert.

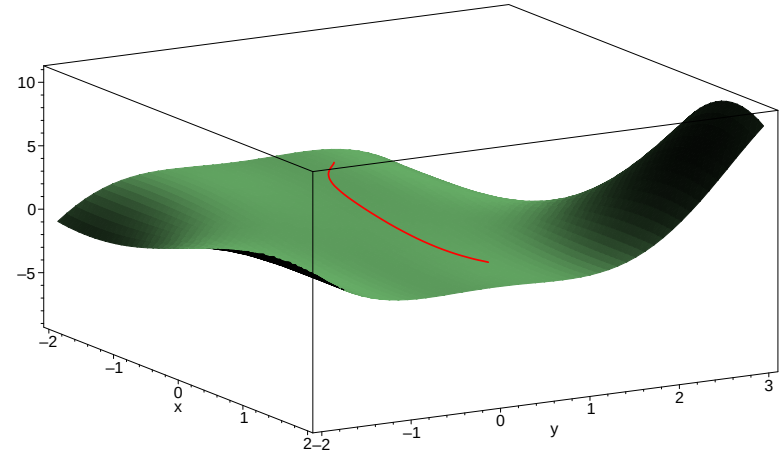
Die folgende Abbildung zeigt ein einfaches Beispiel für einen mit der Gradientenmethode zurückgelegten Weg; hier wurde in jedem Schritt

$$\begin{pmatrix} h_i \\ k_i \end{pmatrix} = 0{,}1 \cdot \nabla f(x_i, y_i)$$

gesetzt. Der Weg geht offensichtlich recht zielstrebig auf das Maximum zu.



Eine Anwendung der Gradientenmethode

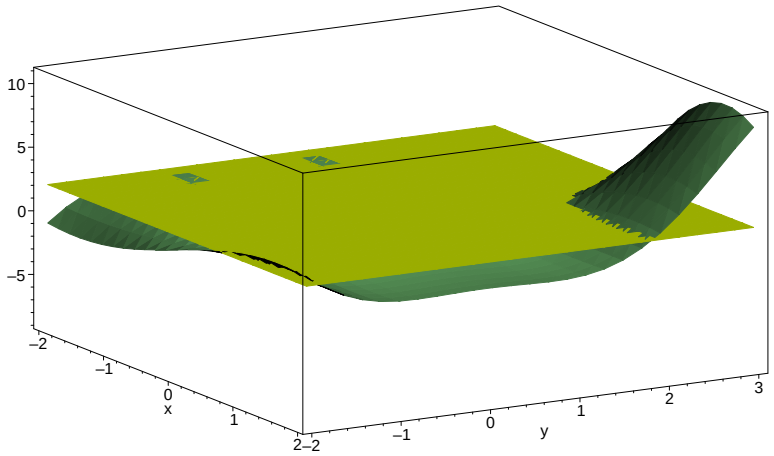


Der Weg aus der vorigen Abbildung aus einem weiteren Blickwinkel

Die darauffolgende Abbildung allerdings zeigt dasselbe Bild in einen etwas größeren Zusammenhang; hier sehen wir, daß unser Streben nach kurzfristigen Gewinnen langfristig wohl doch nicht so erfolgreich war: Wenn wir vom Startpunkt aus nach rechts in die kleine Mulde abgestiegen wären, hätten wir auf dem gegenüberliegenden Hang deutlich größere Funktionswerte erreicht als im lokalen Maximum, in dem wir schließlich gelandet sind. Dies ist ein grundsätzliches Problem von Gradientenverfahren: Falls wir in der Nähe des (absoluten) Optimums starten, führen sie schnell und zuverlässig zum Ziel, ansonsten aber ist die Gefahr sehr groß, daß wir in einem nur lokalen Optimum steckenbleibt.

Um von dort wieder weiterzukommen, gibt es verschiedene Strategien. Eine anschaulich recht klare ist die sogenannte „Tunnelung“. Der Name

entstand aus der Betrachtung von Minimisierungsproblemen; nehmen wir also an, wir wollen das Minimum der Funktion $f(x, y)$ in einem gewissen Bereich finden und ein Gradientenverfahren hat uns in einen Punkt x_M geführt, von dem aus es nicht mehr weiterkommt. Um zu sehen, ob $z_M = f(x_M)$ wirklich der kleinste Wert ist, den f im betrachteten Bereich annehmen kann, versuchen wir, eine weitere Lösung der Gleichung $f(x) = z_M$ zu finden.

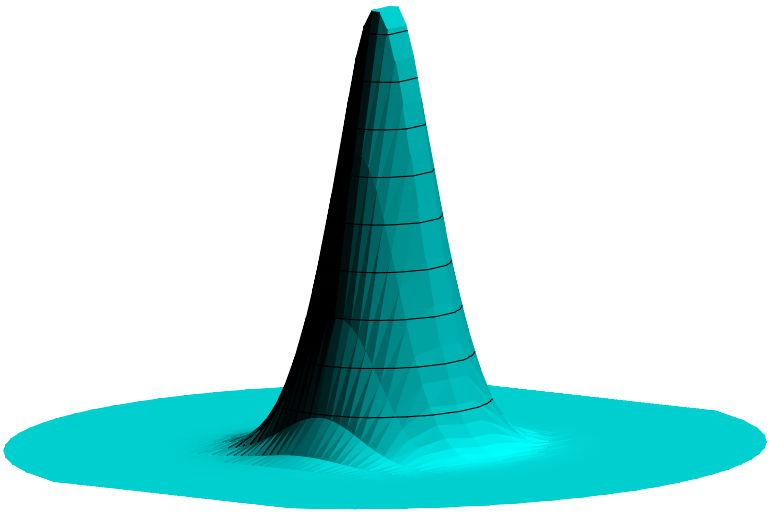


„Tunnelung“ für Maxima

Dafür gibt es eine ganze Reihe numerischer Verfahren, z.B. das Verfahren von NEWTON-RAPHSON, mit denen sich zumindest ein solcher Punkt leicht finden läßt. Leider könnte dieser Punkt unser Ausgangspunkt x_M sein; deshalb sucht man tatsächlich nicht nach Lösungen der Gleichung $f(x) = z_M$, sondern nach Lösungen einer leicht abgewandelten Gleichung der Form $\widetilde{f}(x) = z_M$, wobei $\widetilde{f}$ dadurch aus f entsteht, daß man die Funktionswerte in der unmittelbaren Umgebung von (x_M, y_M) stark anhebt, um so das dortige Minimum zum Verschwinden zu bringen. Dazu kann man beispielsweise eine Funktion der Form

$$G(x, y) = ae^{\frac{(x - x_M)^2 + (y - y_M)^2}{b}}$$

mit geeigneten Parametern a, b wählen, wie sie in der nächsten Abbildung zu sehen ist, und $\widetilde{f}(x, y) = f(x, y) + G(x, y)$ setzen.



$G(x, y) = e^{-3(x^2+y^2)}$

Dies bringt das Minimum im Punkt M zum Verschwinden und verändert die Funktion praktisch nicht, wenn man nur hinreichend weit entfernt ist von M. (Je kleiner b ist, umso lokalisierter ist die Veränderung.) Eine Lösung der Gleichung

$$\widetilde{f}(x) = z_M \ ,$$

so es eine gibt, liegt also nicht in der unmittelbaren Umgebung von x_M und ist daher ein guter Ausgangspunkt, um dort die Gradientenmethode noch einmal zu starten bis zum nächsten lokalen Minimum und so weiter. Sobald die Gleichung nicht mehr lösbar ist, können wir ziemlich sicher sein, daß z_M das globale Minimum ist – es sei denn, wir hätten die Parameter a und b sehr dumm gewählt.

Im obigen Beispiel geht es nicht um ein Minimum, sondern um ein Maximum, da die Suche danach graphisch besser darstellbar ist. Also graben wir auch keinen Tunnel, sondern spannen ein Hochseil, das irgendwo auf der eingezeichneten Ebenen liegt und uns vom erreichten Zwischenhoch zur Startposition für einen weiteren Anstieg bringt. (Tatsächlich ist die Ebene etwas zu tief eingezeichnet, damit man das alte Maximum noch erkennen kann; das Seil muß also etwas höher hängen.)

Eine weitere Idee zur Vermeidung von Zwischenhochs kommt aus der Physik: Ein Gas erreicht seinen Zustand minimaler Energie dann, wenn die Bewegungsenergie $\frac{1}{2}mv^2$ eines jeden Teilchens gleich Null ist, wenn sich also nichts mehr bewegt. Dies geschieht aber höchstens am absoluten Nullpunkt; bei positiven Temperaturen werden die meisten

Teilchen positive kinetische Energie haben. Nach LUDWIG BOLTZMANN ist dabei die Wahrscheinlichkeit dafür, daß ein Teilchen die Energie $E = \frac{1}{2}mv^2$ hat, bei Temperatur T proportional zu

$$e^{-\frac{E}{kT}},$$

mit einer Konstanten $k \approx 1{,}38066 \cdot 10^{-23} J/K$, die heute als BOLTZMANN-Konstante bezeichnet wird.

LUDWIG BOLTZMANN (1844–1906) wuchs auf und studierte in Wien; danach lehrte er in Graz, Heidelberg, Berlin, Graz, Wien, Graz, Wien, Leipzig und Wien. Er war Professor für Theoretische Physik, für Mathematik und für Experimentalphysik. Auf seiner letzten Stelle in Wien hielt er eine so erfolgreiche Philosophievorlesung, daß ihn Kaiser Franz Josef in den Palast einlud. Am bekanntesten ist er für die Begründung der statistischen Mechanik, einer damals sehr umstrittene Theorie. Ob die damit verbundenen Anfeindungen zu seinem Selbstmord führten, ist unbekannt.

Die folgende Abbildung zeigt für verschiedene Werte von kT die Graphen der Funktion $e^{-E/kT}$; wie man sieht; erwartungsgemäß sind diese für große Werte von kT sehr flach, während sie für kleine Temperaturen rechts schnell gegen Null gehen. Bei der *simulierten Abkühlung* ahmt man dies nach, indem man mit einer hohen Temperatur startet und der Richtung, in der man weitergeht, einer dieser Temperatur entsprechende Freiheit läßt.

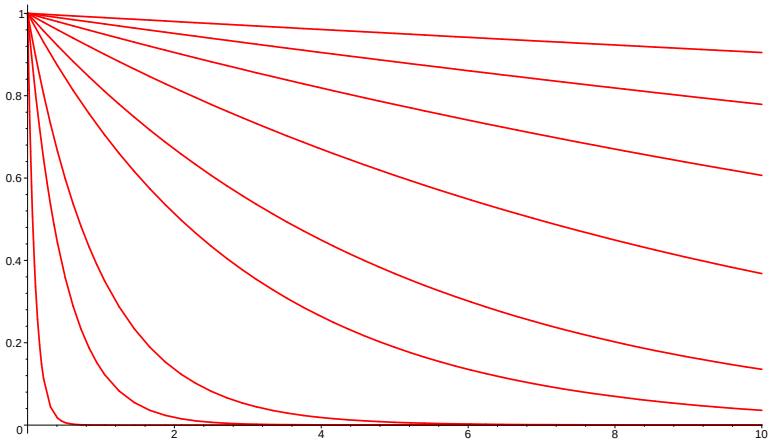
Man geht also nicht mehr unbedingt in Richtung des Gradienten, sondern geht zufällig in eine von endlich vielen vorgegebenen Richtungen. Die Wahrscheinlichkeit für den Richtungsvektor h_j soll dabei analog zur BOLTZMANN-Verteilung festgelegt werden, d.h. wir ordnen ihm eine „Energie“ $E_j = \pm\big(f(x + h_j) - f(x, y)\big)$ zu (positiv bei der Suche nach einem Minimum, negativ bei der Suche nach einem Maximum) und die Wahrscheinlichkeit dafür, daß wir in Richtung h_j gehen, soll proportional sein zu $e^{-E_j/kT}$. Sie ist also, falls N Richtungen zur Verfügung stehen, gleich

$$p_j \underset{\text{def}}{=} e^{\frac{-E_j/kT}{\sum_{\ell=1}^{N} e^{-E_\ell/kT}}}.$$

Zur Wahl einer Richtung erzeugen wir uns eine Zufallszahl $Z \in [0, 1]$ und gehen in Richtung h_j, wenn

$$\sum_{\ell=1}^{j-1} p_\ell < Z \leq \sum_{\ell=1}^{j} p_\ell$$

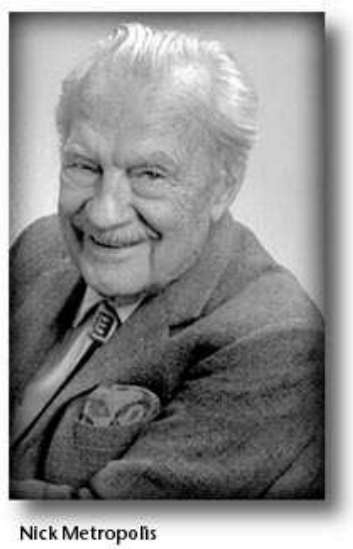
ist. (Die Frage, wie lang die Richtungsvektoren im wievielten Schritt sein sollen, wollen wir hier ausklammern.)



$e^{-E/kT}$ für kT = 0,1, 0,5, 1, 3, 5, 10, 20, 40, 100

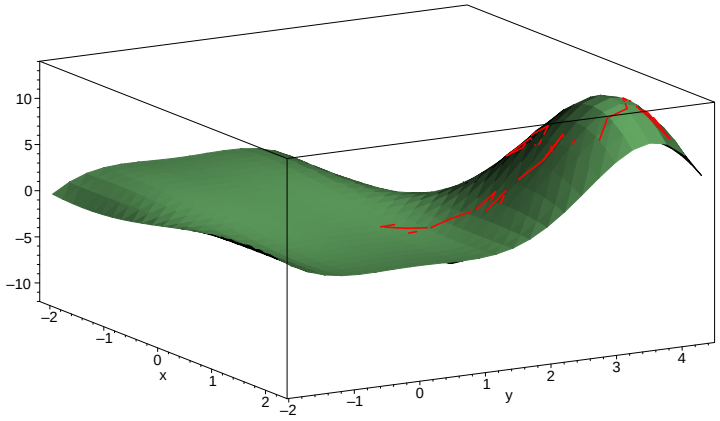
Bei hohen Temperaturen ist damit die Richtung fast vollständig zufallsbedingt gewählt, während in der Nähe des absoluten Nullpunkts praktisch nur noch die optimale Richtung eine Chance hat. Falls wir bei hoher Temperatur in einem Zwischenextremum landen, sorgt dies mit sehr hoher Wahrscheinlichkeit dafür, daß wir dort nicht steckenbleiben.

Am Ende wollen wir allerdings zum absoluten Optimum kommen, d.h. wir müssen die Temperatur im Verlauf der Rechnung immer weiter senken – daher der Name *simulated annealing* = simulierte Abkühlung. Bei der Anwendung auf Optimierungsprobleme bezeichnet man diese Vorgehensweise als den METROPOLIS-Algorithmus. In welcher Weise man die Temperatur am besten senkt, ist immer noch ein Gebiet aktiver Forschung. Man kann zeigen, daß man statistisch betrachtet praktisch immer im Optimum landet, wenn man mit einer hinreichend hohen Ausgangstemperatur T_1 startet und im r-ten Schritt mit Temperatur $T_1/\log(r + 1)$ arbeitet, aber bei einer derart langsamen Abkühlung braucht der Algorithmus viel zu lange, um ans Ziel zu kommen.

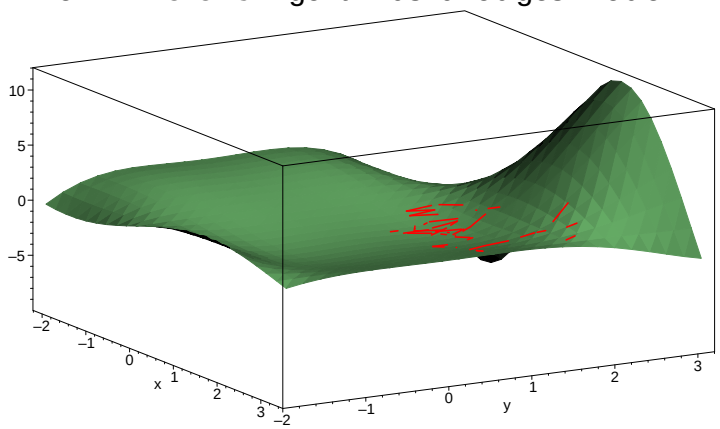

Nick Metropolis

NICHOLAS METROPOLIS (1915–1999) wuchs auf in Chicago, wo er Physik studierte und 1941 promovierte. Seit 1943 arbeitete er, unterbrochen durch Professuren an der Universität Chicago von 1945–1948 und 1957–1965, in den Los Alamos Labaratories, die ihn im Nachruf als *giant of mathematics and one of the founders of the Information Age* bezeichneten. Sein Ruhm als Mathematiker beruht vor allem auf den von ihm entwickelten Anwendungen statistischer Verfahren auf eine Vielzahl mathematischer Probleme; zum Pionier des Informationszeitalter macht ihn u. a., daß er einer der ersten Anwender des ersten elektronischen Computers ENIAC war, dessen Nachfolger MANIAC baute und an der Universität Chicago das Institute for Computer Research gründete und bis 1965 leitete.

Die beiden folgenden Abbildungen zeigen, wie sich der Algorithmus bei zwei verschiedenen Folgen von Zufallszahlen verhält bei einer Abkühlungsregel, die im r-ten Schritt mit Temperatur T_1/r arbeitet:



Der METROPOLIS-Algorithmus für obiges Problem



Dito mit anderen Zufallszahlen

Im ersten Beispiel funktioniert alles sehr gut, im zweiten dagegen bleibt die Kurve ziemlich lange im Tal hängen, kommt aber immerhin in eine gute Startposition für weitere Iterationen. Oft wird es ohnehin am besten sein, nach hinreichend vielen METROPOLIS-Schritten ein gewöhnliches Gradientenverfahren zu starten.

Zusammenfassend läßt sich sagen, daß der METROPOLIS-Algorithmus und verwandte Verfahren (die sogenannten Monte-Carlo-Methoden) sehr nützliche Hilfsmittel zur Optimierung sind, falls man so gut wie nichts über die zu optimierende Funktion weiß. Sie funktionieren nicht nur bei kontinuierlichen Problemen, wie den hier betrachteten, sondern auch für diskrete und kombinatorische Optimierungsprobleme, haben aber den Nachteil, daß sie kein Optimum garantieren können: Selbst wenn man eines erreicht hat, kann die Methode dies nicht erkennen. (Es gibt alternative numerische Methoden, die das können.)

Wie schon diese sehr kleine Auswahl von Optimierungsverfahren zeigt, ist nichtlineare Optimierung ein sehr weites Feld, von dem wir hier nur einen winzigen Ausschnitt betrachten konnten. Dieser Ausschnitt bestand nicht aus den für die Praxis wichtigsten Verfahren, sondern aus denen, die sich am besten in den Stoff der Vorlesung einordnen. Sie sind zwar (in Kombination mit anderen Verfahren) die Grundbausteine, aus denen sich die meisten praktisch relevanten Verfahren zusammengesetzen, aber für die vielen kleinen Abwandlungen, die dazu führen, daß man ein Problem wirklich effizient lösen kann, müßten wir deutlich mehr Zeit aufwenden, als hier zur Verfügung steht. Interessenten seien auf Spezialvorlesungen über Optimierung verwiesen.

§3: Höhere Ableitungen

Im Eindimensionalen ist die Ableitung einer Funktion $f\colon D \to \mathbb{R}$ wieder eine Funktion $D \to R$; falls sie differenzierbar ist, bezeichnen wir ihre Ableitung als die zweite Ableitung von f, und so weiter. Für eine Funktion $f\colon D \to \mathbb{R}$ von n Veränderlichen ist die Ableitung jedoch eine Funktion $D \to \mathbb{R}^n$, also etwas komplizierteres als die Ausgangsfunktion. Dies macht den Umgang mit höheren Ableitungen im Mehrdimensionalen schwieriger. Wir beschränken uns daher zunächst auf die zweite Ableitung einer reellwertigen Funktion.

a) Die Hesse-Matrix

Wir lassen uns vom eindimensionalen Fall leiten: Die zweite Ableitung ist die Ableitung der Ableitung.

Die Ableitung einer differenzierbaren Funktion $f\colon D \to \mathbb{R}$ mit $D \subseteq \mathbb{R}^n$ ist der Gradient von f, also die Abbildung

$$\nabla f\colon D \to \mathbb{R}^n; \quad x \mapsto \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n} \right) .$$

Im vorigen Paragraphen haben wir auch solche Funktionen differenziert; ihre Ableitung war eine Matrix.

Definition: Wenn der Gradient von $f\colon D \to \mathbb{R}$ differenzierbar ist, bezeichnen wir seine JACOBI-Matrix als die HESSE-Matrix

$$H_f(x) = J_{\nabla f}(x)$$

von f im Punkt x.

Die HESSE-Matrix ist somit stets quadratisch, denn die Anzahl der Komponenten des Gradienten ist gleich der Anzahl der Variablen.

Genau wie der Gradient bei gutartigen Funktionen durch die partiellen Ableitungen berechnet werden kann, sollte auch sie durch Differentiationsverfahren aus der Analysis einer Veränderlichen berechenbar sein. Das Hilfsmittel dazu sind die zweiten partiellen Ableitungen.

LUDWIG OTTO HESSE (1811–1874) wurde in Königsberg geboren und unterrichtete zunächst Physik und Chemie am dortigen Gymnasium. 1840 bekam er eine Stelle als Mathematiker an der dortigen Universität, von 1856 bis 1868 war er Professor in Heidelberg, danach in München. Aus der Schule ist er wohl vor allem durch die HESSEsche Normalenform der Ebenengleichung bekannt; der Schwerpunkt seiner Forschungen lag allerdings auf dem Gebiet der Invariantentheorie und der algebraischen Funktionen. Auch die HESSE-Matrix führte er 1842 in einer Arbeit über Invarianten von kubischen und biquadratischen Kurven ein.

Für eine in ganz D partiell differenzierbare Funktion $f\colon D \to \mathbb{R}$ ist auch jede partielle Ableitung f_{x_i} wieder eine Funktion von D nach $\mathbb{R}$, und

auch diese kann wieder partiell differenzierbar sein. Falls ja, bezeichnen wir die partielle Ableitung von f_{x_i} nach x_j als zweite partielle Ableitung

$$f_{x_i x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i} \underset{\text{def}}{=} \frac{\partial}{\partial x_j}\left(\frac{\partial f}{\partial x_i}\right)$$

von f nach x_i und x_j. Im Fall $i = j$ schreiben wir kurz

$$f_{x_i x_i} = \frac{\partial^2 f}{\partial x_i^2} .$$

Analog lassen sich auch höhere partielle Ableitungen einführen durch die Definition

$$f_{x_{i_1} x_{i_2} \ldots x_{i_k}} \underset{\text{def}}{=} \frac{\partial^k f}{\partial x_{i_k} \ldots \partial x_{i_2} \partial x_{i_1}} \underset{\text{def}}{=} \frac{\partial}{\partial x_{i_k}} \frac{\partial}{\partial x_{i_{k-1}}} \cdots \frac{\partial}{\partial x_{i_2}} \frac{\partial f}{\partial x_{i_1}} .$$

Wie wir oben gesehen haben, ist Gradient einer Funktion f gleich dem Vektor der partiellen Ableitungen, falls diese allesamt existieren und stetig sind; die JACOBI-Matrix ist der entsprechende Zeilenvektor. Falls auch die zweiten partiellen Ableitungen allesamt existieren und stetig sind, zeigt dasselbe Lemma, daß deren Ableitungen die Zeilenvektoren

$$\left(\frac{\partial^2 f}{\partial x_i \partial x_1}, \ldots, \frac{\partial^2 f}{\partial x_i \partial x_n}\right)$$

sind, d.h.

Lemma: Falls alle ersten und zweiten partiellen Ableitungen von $f\colon D \to \mathbb{R}$ existieren und stetig sind, ist die HESSE-Matrix von f gleich der $n \times n$-Matrix mit Einträgen $\frac{\partial^2 f}{\partial x_i \partial x_j}$.

■

Als erstes Beispiel können wir etwa die zweiten partiellen Ableitungen der Funktion

$$f(x, y) = x^4 + 2x^3y + 3x^2y^2 + 4xy^3 + 5y^4$$

berechnen: Die partielle Ableitung nach x ist

$$f_x(x, y) = 4x^3 + 6x^2y + 6xy^2 + 4y^3 ,$$

also ist

$$f_{xx}(x, y) = \frac{\partial^2 f}{\partial x^2}(x, y) = 12x^2 + 12xy + 6y^2 \quad \text{und}$$

$$f_{xy}(x,y) = \frac{\partial^2 f}{\partial y \partial x}(x,y) = 6x^2 + 12xy + 12y^2 \,.$$

Entsprechend ist

$$f_y(x,y) = 2x^3 + 6x^2y + 12xy^2 + 20y^3 \,,$$

also

$$f_{yx}(x,y) = \frac{\partial^2 f}{\partial x \partial y}(x,y) = 6x^2 + 12xy + 12y^2 \quad \text{und}$$
$$f_{yy}(x,y) = \frac{\partial^2 f}{\partial y^2}(x,y) = 6x^2 + 24xy + 60y^2 \,.$$

Damit ist

$$H_f(x,y) = \begin{pmatrix} 12x^2 + 12xy + 6y^2 & 6x^2 + 12xy + 12y^2 \\ 6x^2 + 12xy + 12y^2 & 6x^2 + 24xy + 60y^2 \end{pmatrix} \,.$$

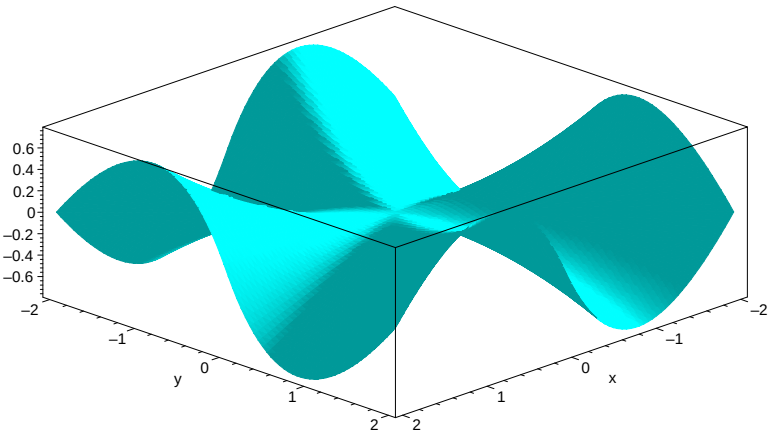
Zumindest in diesem Fall ist dies eine symmetrische Matrix, d.h.

$$f_{xy} = f_{yx} \,.$$

Diese Formel gilt, wie wir gleich sehen werden, *fast* immer; in der Tat galt sie für die Mathematiker des 18. Jahrhunderts wie NICOLAUS I. BERNOULLI, der 1719 darüber schrieb, LEONARD EULER (1730), JOSEPH-LOUIS LAGRANGE (1772) und viele andere als selbstverständlich. Erst im 19. Jahrhundert, als sich ein präziser Funktionsbegriff durchzusetzen begann, wurde erkannt, daß Voraussetzungen notwendig sind. Diese waren zu Beginn des Jahrhunderts zunächst unnötig stark; erst 1873 fand HERMANN AMANDUS SCHWARZ in seiner Arbeit *Über ein System voneinander unabhängiger Voraussetzungen zum Beweis des Satzes* $\frac{\partial}{\partial y}\left(\frac{\partial f(x,y)}{\partial x}\right) = \frac{\partial}{\partial x}\left(\frac{\partial f(x,y)}{\partial y}\right)$, was wirklich notwendig ist.

Als Gegenbeispiel betrachtet er die in der folgenden Abbildung dargestellte Funktion

$$f(x,y) = \begin{cases} y^2 \arctan \frac{x}{y} - x^2 \arctan \frac{y}{x} & \text{für } (x,y) \neq (0,0) \\ 0 & \text{für } (x,y = (0,0) \end{cases} \,.$$



Ein Gegenbeispiel zum Vertauschungssatz

Auch wenn es auf den ersten Blick nicht so aussieht, ist diese Funktion auch für $y = 0$ und $x \neq 0$ definiert: Der Bruch x/y ist dann zwar nicht definiert, aber da der Arkustangens nur Werte zwischen $-\pi/2$ und $\pi/2$ annimmt, existiert für jedes $x \neq 0$ der Grenzwert $\lim_{y\to 0} y^2 \arctan \frac{x}{y}$ und verschwindet. Entsprechend ist auch $\lim_{x\to 0} x^2 \arctan \frac{y}{x} = 0$ für alle $y \neq 0$, und wir wollen die obige Formel so interpretieren, daß $f(0,y) = f(x,0) = 0$ sein soll für alle $x, y \neq 0$ und, nach Definition, natürlich auch für $x = y = 0$. Man überlegt sich leicht, daß die so definierte Funktion stetig ist auf ganz $\mathbb{R}^2$.

Die Berechnung der ersten partiellen Ableitungen ist etwas umfangreich, jedoch läßt sich das Ergebnis deutlich vereinfachen: Wir erhalten

$$f_x(x,y) = \begin{cases} y - 2x \arctan \frac{y}{x} & \text{für } (x,y) \neq (0,0) \\ 0 & \text{für } (x,y) = (0,0) \end{cases}$$

und

$$f_y(x,y) = \begin{cases} -x + 2y \arctan \frac{x}{y} & \text{für } (x,y) \neq (0,0) \\ 0 & \text{für } (x,y) = (0,0) \end{cases},$$

wobei die Division durch Null beim Argument des Arkustangens wegen des Faktors vor dem Arkustangens wieder wie oben interpretiert werden soll, wir haben also

$$f_x(0,y) = y \quad \text{und} \quad f_y(x,0) = -x\,.$$

Diese Funktionen können wir problemlos differenzieren; wir erhalten

$$f_{xy}(0,y) = +1 \quad \text{und} \quad f_{yx}(x,0) = -1\,.$$

Im Nullpunkt ist somit $f_{xy}(0,0) = +1 \neq -1 = f_{yx}(0,0)$.

Für Punkte $(x, y) \neq (0, 0)$ rechnet man leicht nach, daß

$$f_{xy}(x, y) = f_{yx}(x, y) = \frac{y^2 - x^2}{y^2 + x^2}$$

ist. Insbesondere ist daher für $x, y \neq 0$

$$f_{xy}(x, 0) = -1, \quad f_{xy}(0, y) = +1 \quad \text{und} \quad f_{xy}(x, x) = 0\,;$$

f_{xy} nimmt also in jeder noch so kleinen Umgebung des Nullpunkts jeden der drei Werte $0, 1$ und -1 (und viele andere) an. Damit kann f_{xy} in $(0, 0)$ nicht stetig sein, genauso wenig wie f_{yx}. Wie SCHWARZ erkannte, ist genau das die fehlende Voraussetzung für die Vertauschbarkeit der partiellen Ableitungen:

Schwarzsches Lemma: $f\colon D \to \mathbb{R}$ sei auf $D \subseteq \mathbb{R}^n$ erklärt, und sowohl die ersten partiellen Ableitungen f_{x_i} als auch die gemischten partiellen Ableitungen $f_{x_i x_j}$ seien stetig auf D. Dann ist

$$f_{x_i x_j}(x) = f_{x_j x_i}(x)$$

für alle $x \in D$ und alle i, j mit $1 \leq i, j \leq n$.

Beweis: Da bei der partiellen Differentiation alle Variablen außer einer als konstant betrachtet werden, können wir uns auf den Fall $n = 2$ beschränken: Wir interessieren uns nur für die beiden Variablen x_i und x_j, die wir als x und y bezeichnen (wenn sie verschieden sind – andernfalls gibt es aber ohnehin nichts zu beweisen), und betrachten alle sonstigen x_k als konstant.

Für den Punkt (x, y) aus D wählen wir dann $h, k \in \mathbb{R}$ so, daß das Quadrat mit den vier Ecken

$$(x, y),\ (x + h, y),\ (x, y + k) \quad \text{und} \quad (x + h, y + k)$$

vollständig in D liegt; dies ist möglich, da wir D als offene Menge vorausgesetzt haben. Nach Voraussetzung existieren die partiellen Ableitungen f_x, f_y, f_{xy} sowie f_{yx} und sind stetig.

Nach Definition ist

$$f_{xy}(x, y) = \lim_{k \to 0} \frac{f_x(x, y + k) - f_x(x, y)}{k}$$

$$= \lim_{k \to 0} \frac{\lim\limits_{h \to 0} \dfrac{f(x + h, y + k) - f(x, y + k)}{h} - \lim\limits_{h \to 0} \dfrac{f(x + h, y) - f(x, y)}{h}}{k}\,.$$

Falls alle Grenzübergänge miteinander vertauschbar sind (was, wie wir im obigen Gegenbeispiel gesehen haben, keineswegs selbstverständlich ist), ist das ein Limes über den Ausdruck

$$\frac{f(x+h,y+k)-f(x,y+k)-f(x+h,y)+f(x,y)}{hk}$$

für $h,k \to 0$; es liegt also nahe, sich diesen Bruch genauer anzuschauen.

Für den Beweis wird es genügen, wenn wir uns auf den Fall $h = k$ beschränken; wir wollen den Ausdruck

$$D(h) = \frac{f(x+h,y+h)-f(x,y+h)-f(x+h,y)+f(x,y)}{h^2}$$

auf zwei Arten ausrechnen:

Zunächst fassen wir, wie oben, die beiden ersten und die beiden letzten Summanden zusammen: Mit der Abkürzung

$$g(y) = \frac{f(x+h,y)-f(x,y)}{h}$$

ist dann

$$D(h) = \frac{g(y+h)-g(y)}{h}\,.$$

Nach dem Mittelwertsatz der Differentialrechnung ist dieser Differenzenquotient gleich dem Differentialquotient $g'(\eta)$ für eine (von h abhängige) Zahl η zwischen y und $y+h$. Somit ist

$$D(h) = g'(\eta) = \frac{f_y(x+h,\eta)-f_y(x,\eta)}{h} = f_{yx}(\xi,\eta)$$

für ein η zwischen x und $x+h$, denn natürlich können wir auch auf diesen Differenzenquotienten den Mittelwertsatz anwenden.

Für die zweite Berechnung fassen wir in $D(h)$ den ersten und den dritten sowie den zweiten und den vierten Term zusammen. Mit der Abkürzung

$$\widetilde{g}(x) = \frac{f(x,y+h)-f(x,y)}{h}$$

ist dann dieses Mal

$$D(h) = \frac{\widetilde{g}(x+h)-\widetilde{g}(y)}{h}\,,$$

und nach dem Mittelwertsatz der Differentialrechnung gibt es dazu ein $\widetilde{\xi}$ zwischen x und $x + h$, so daß dies gleich $\widetilde{g}'(\widetilde{\xi})$ ist. Also ist

$$D(h) = \widetilde{g}'(\widetilde{\xi}) = \frac{f_x(\widetilde{\xi}, y + h) - f_x(\widetilde{\xi}, y)}{h} = f_{xy}(\widetilde{\xi}, \widetilde{\eta})$$

für eine Zahl $\widetilde{\eta}$ zwischen y und $y + h$. Somit ist

$$D(h) = f_{yx}(\xi, \eta) = f_{xy}(\widetilde{\xi}, \widetilde{\eta})\,.$$

Lassen wir nun h gegen Null gehen, konvergieren ξ und $\widetilde{\xi}$ gegen x und η wie auch $\widetilde{\eta}$ gegen y. Wegen der vorausgesetzten Stetigkeit der zweiten partiellen Ableitungen konvergiert daher $f_{yx}(\xi, \eta)$ gegen $f_{yx}(x, y)$ und $f_{xy}(\widetilde{\xi}, \widetilde{\eta})$ gegen $f_{xy}(x, y)$, d.h. der Grenzwert existiert und

$$D(0) = f_{yx}(x, y) = f_{xy}(x, y)\,.$$

Damit ist das Lemma bewiesen. ∎

Der deutsche Mathematiker KARL HERMAN AMANDUS SCHWARZ (1843–1921) beschäftigte sich hauptsächlich mit konformen Abbildungen und mit sogenannten Minimalflächen, d.h. Flächen mit vorgegebenen Eigenschaften, deren Flächeninhalt minimal ist. Im Rahmen einer entsprechenden Arbeit für die WEIERSTRASS-Festschrift von 1885 (im Falle eines durch Doppelintegrale definierten Skalarprodukts) bewies er die CAUCHY-SCHWARZsche Ungleichung, die CAUCHY bereits 1821 für endlichdimensionale Vektorräume bewiesen hatte. SCHWARZ lehrte nacheinander in Halle, Zürich, Göttingen und Berlin.

Tatsächlich bewies SCHWARZ das obige Lemma (für $n = 2$) unter einer etwas schwächeren Voraussetzung: Es reicht, wenn *eine* der partiellen Ableitungen f_{xy} oder f_{yx} existiert und stetig ist. Am Beweis ändert sich wenig; falls etwa über die Ableitung f_{yx} nichts vorausgesetzt ist, muß man die Existenz aller damit zusammenhängenden Grenzwerte explizit durch Abschätzungen nachweisen und daraus nachträglich die Existenz und Stetigkeit von f_{yx} folgern. Ein Leser, der seine *Analysis I* noch nicht ganz vergessen hat, sollte dies auf etwa einer Seite tun können. Für Anwendungen ist die SCHWARZsche Formulierung etwas nützlicher als die obige, denn wenn man beispielsweise f_{xy} berechnet und seine Stetigkeit nachgewiesen hat, folgt automatisch, daß auch f_{yx} existiert und gleich f_{xy} ist. Für uns wird das keine sehr große Rolle spielen, denn bei den meisten uns interessierenden Funktionen wird die Existenz und Stetigkeit der partiellen Ableitungen klar sein; lediglich ihre Berechnung wird im allgemeinen mit Arbeit verbunden sein.

Ein analoger Satz zum SCHWARZschen Lemma gilt auch für höhere partielle Ableitungen; für k-fache Ableitungen müssen wir natürlich voraussetzen, daß alle partiellen Ableitungen bis zu den k-fachen existieren und stetig sind (wobei diese Voraussetzung wieder streng genommen nicht für alle k-fachen wirklich notwendig ist).

Definition: Für eine offene Teilmenge $D \subseteq \mathbb{R}^n$ bezeichne $\mathcal{C}^k(D,\mathbb{R})$ die Menge aller Funktionen $f\colon D \to \mathbb{R}$, deren sämtliche partielle Ableitungen bis zu den k-ten existieren und stetig sind. Für $k = 0$ bezeichnen wir mit $\mathcal{C}^0(D,\mathbb{R})$ einfach die Menge aller stetiger Funktionen $D \to \mathbb{R}$.

Man überlegt sich sofort, daß $\mathcal{C}^k(D,\mathbb{R})$ ein $\mathbb{R}$-Vektorraum ist, und es ist auch nicht schwer einzusehen, daß die Funktionen aus $\mathcal{C}^k(D,\mathbb{R})$ alle die Eigenschaften haben, die man sich bei der Betrachtung von k-ten Ableitungen wünscht:

Erstens ist die Berechnung einer k-ten partiellen Ableitung von der Reihenfolge der partiellen Differentiationen unabhängig: Wie aus der *Linearen Algebra* bekannt sein sollte, kann jede Permutation als Produkt von Transpositionen geschrieben werden; es genügt also zu zeigen, daß man die Reihenfolge zweiter partieller Differentiationen vertauschen kann. Eine Transposition $(i_r\ i_{r+k})$ wiederum läßt sich gemäß

$$(i_r\ i_{r+k}) = (i_r\ i_{r+1})\cdots(i_{r+k-1}\ i_{r+k})(i_{r+k-2}\ i_{r+k-1})\cdots(i_r\ i_{r+1})$$

als Produkt von Transpositionen benachbarter Elemente schreiben, und für eine solche Transposition ist

$$\frac{\partial}{\partial x_{i_1}}\cdots\frac{\partial}{\partial x_{i_r}}\frac{\partial}{\partial x_{i_{r+1}}}\cdots\frac{\partial f}{\partial x_{i_k}} = \frac{\partial}{\partial x_{i_1}}\cdots\frac{\partial}{\partial x_{i_{r+1}}}\frac{\partial}{\partial x_{i_r}}\cdots\frac{\partial f}{\partial x_{i_k}}\,.$$

Für $f \in \mathcal{C}^k(D,\mathbb{R})$ ist sichergestellt, daß $\frac{\partial}{\partial x_{i_{r+2}}}\cdots\frac{\partial f}{\partial x_{i_k}}$ in $\mathcal{C}^2(D,\mathbb{R})$ liegt; nach obigem Lemma ist daher

$$\frac{\partial}{\partial x_{i_r}}\frac{\partial}{\partial x_{i_{r+1}}}\left(\frac{\partial}{\partial x_{i_{r+2}}}\cdots\frac{\partial f}{\partial x_{i_k}}\right) = \frac{\partial}{\partial x_{i_{r+1}}}\frac{\partial}{\partial x_{i_r}}\left(\frac{\partial}{\partial x_{i_{r+2}}}\cdots\frac{\partial f}{\partial x_{i_k}}\right)\,.$$

Differenziert man hier beide Seiten noch partiell nach x_{i_1} bis $x_{i_{r-1}}$, ändert dies natürlich nichts an der Gleichheit.

Zweitens besagt das vorletzte Lemma, daß eine Funktion $f \in \mathcal{C}^1(D,\mathbb{R})$ differenzierbar ist. Eine nahe liegende Verallgemeinerung des dortigen

Beweises, bei der man anstelle von linearen Approximationen solche höherer Ordnung betrachtet, zeigt, daß eine Funktion $f \in \mathcal{C}^2(D, \mathbb{R})$ zweifach differenzierbar ist, und daß entsprechend eine Funktion f aus $\mathcal{C}^k(D, \mathbb{R})$ bis auf einen Fehler der Größenordnung $o\big(|h|^k\big)$ durch ein Polynom k-ten Grades approximiert werden kann. Wie das im einzelnen aussieht, wollen wir uns im nächsten Abschnitt genauer anschauen.

b) Taylor-Polynome

Im letzten Semester haben wir die höheren Ableitungen einer Funktion dazu benutzt, um sie nicht nur durch eine lineare Funktion, sondern durch ein Polynom höheren Grades anzunähern. Der wesentliche Satz über TAYLOR-Polynome war der folgende:

Satz: $f\colon (a, b) \to \mathbb{R}$ sei stetig und mindestens $(k + 1)$-fach stetig differenzierbar auf dem Intervall $(a, b) \subseteq \mathbb{R}$. Dann gilt für jedes x aus (a, b) und jedes $h \in \mathbb{R}$ mit $x + h \in (a, b)$ die Formel

$$\begin{aligned} f(x+h) &= f(x) + hf'(x) + \frac{h^2}{2} f''(x) + \cdots + \frac{h^k}{k!} f^{(k)}(x) + R_{k+1}(x,h) \\ &= \sum_{i=0}^{k} \frac{h^i}{i!} f^{(i)}(x) + R_{k+1}(x,h) \end{aligned}$$

mit einem Restglied $R_{k+1} = O(h^{k+1})$. Dieses kann beispielsweise dargestellt werden als

$$R_{k+1}(x,h) = \frac{h^{k+1}}{(k+1)!} f^{(k+1)}(x + \eta h)$$

mit einer reellen Zahl η zwischen 0 und 1.

Um daraus einen auch für Funktionen mehrerer Veränderlichen nützlichen Satz zu machen, verwenden wir wieder Richtungsableitungen. Da wir eine Funktion $f\colon \mathbb{R}^n \to \mathbb{R}^m$ problemlos komponentenweise behandeln können, genügt es, den Fall $m = 1$ zu betrachten.

Die Richtungsableitung einer Funktion $f\colon \mathbb{R}^n \to \mathbb{R}$ in Richtung v ist nach Definition die Ableitung der Funktion einer Veränderlichen

$$g\colon \begin{cases} (-a, a) & \to \mathbb{R} \\ t & \mapsto f(x + tv) \end{cases}$$

mit geeignet gewähltem $a \in \mathbb{R}_+$ nach t für $t = 0$; wir können sie berechnen als Skalarprodukt des Gradienten mit dem Vektor v.

Natürlich können wir g nicht nur einmal ableiten; für $f \in \mathcal{C}^{k+1}(D, \mathbb{R})$ existiert das TAYLOR-Polynom k-ten Grades und

$$g(t) = g(0) + tg'(0) + \frac{t^2}{2}g''(0) + \cdots + \frac{t^k}{k!}g^{(k)}(0) + O(t^{k+1}) .$$

Schreiben wir die Richtungsableitung in Richtung v wieder als

$$\partial_v f(x) = \langle \nabla f(x), v \rangle = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x) \cdot v_i ,$$

so wird dies zu

$$f(x + tv) = f(x) + t\partial_v f(x) + \cdots + \frac{t^k}{k!}\partial_v^k f(x) + O(t^{k+1}) .$$

Da wir $\|v\|$ hier als eine Konstante betrachten, können wir $O(t^{k+1}$ auch schreiben als $O\big((t\,\|v\|)^{k+1}$ und damit insbesondere $p\big((t\,\|v\|)^k$; speziell für $t = 1$ erhalten wir die kompakte Schreibweise der TAYLOR-Formel, nämlich

$$f(x + v) = f(x) + \partial_v f(x) + \cdots + \frac{1}{k!}\partial_v^k f(x) + o\big(\|v\|^k\big) .$$

∂_v^k steht hierbei natürlich für die k-fache Anwendung des Operators ∂_v.

Wenn wir das konkret ausrechnen müssen, verschwindet die Kompaktheit allerdings schnell: Mit

$$v = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}$$

ist $\partial_v f(x) = \langle \nabla f(x), v \rangle = \sum_{i=1}^{n} v_i f_{x_i}(x)$ und dementsprechend

$$\partial_v^2 f(x) = \partial_v\big(\partial_v f(x)\big) = \partial_v \left(\sum_{i=1}^{n} v_i f_{x_i} \right) = \sum_{j=1}^{n} v_j \frac{\partial}{\partial x_j} \left(\sum_{i=1}^{n} v_i f_{x_i} \right)$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} v_i v_j f_{x_i x_j} ,$$

wir haben also schon eine Summe mit n^2 Termen.

Mit Hilfe der linearen Algebra können wir diese Summe noch relativ kurz schreiben: Da der (i, j)-Eintrag der HESSE-Matrix $H_f(x)$ gerade

gleich die partielle Ableitung $f_{x_i x_j}$ ist, rechnet man leicht nach, daß

$$v^T H_f(x) v = (v_1, \ldots, v_n) H_f(x) \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} = \sum_{i=1}^{n} \sum_{j=1}^{n} v_i v_j f_{x_i x_j}$$

ist. Für höhere Ableitungen geht so etwas nicht mehr: Völlig analog zur obigen Rechnung überzeugt man sich leicht davon, daß

$$\partial_v^3 f(x) = \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} v_i v_j v_k f_{x_i x_j x_k}$$

ist, und diese Summe aus n^3 Summanden läßt sich nicht mehr mit Matrix-Vektor-Produkten darstellen: Die dritte Ableitung ist gegeben durch einen sogenannten *Tensor dritter Stufe*, d.h. ein dreidimensionales würfelförmiges Zahlenschema, und mit jeder weiteren Ableitung steigt die Dimension um eins an.

Für die Diskussion im nächsten Abschnitt reicht uns glücklicherweise die Formel

$$f(x+h) = f(x) + \langle \nabla f(x), h \rangle + \frac{1}{2} h^T H_f(x) h + o(\|h\|^2) \,.$$

c) Höhere Ableitungen und lokale Extrema

Wenn grad $f(x_0)$ verschwindeti und f mindestens zweimal stetig differenzierbar ist, hängt also das Verhalten von f in der Umgebung von x_0 ab von der quadratischen Form $h \mapsto h^T H_f(x_0) h$, wobei H_f nach dem SCHWARZschem Lemma eine symmetrische Matrix ist.

Definition: *a)* Eine symmetrische Matrix $A \in \mathbb{R}^{n \times n}$ heißt *positiv definit*, wenn für alle Vektoren $v \neq 0$ aus $\mathbb{R}^n$ gilt: $v^T A v > 0$.
b) A heißt *negativ definit*, wenn für alle $v \neq 0$ aus $\mathbb{R}^n$ gilt: $v^T A v < 0$.
c) A heißt *indefinit*, wenn es Vektoren $v, w \in \mathbb{R}^n$ gibt, so daß gilt: $v^T A v > 0$, aber $w^T A w < 0$.

Mit dieser Terminologie ist das folgende Lemma klar:

Satz: Wenn die differenzierbare Funktion $f \in \mathcal{C}^1(D, \mathbb{R})$ im Punkt $x_0 \in D$ ein lokales Extremum hat, ist dort ihr Gradient gleich dem Nullvektor.

Falls umgekehrt für $f \in \mathcal{C}^2(D, \mathbb{R})$ der Gradient im Punkt $x \in D$ verschwindet, gilt:
a) Falls die HESSE-Matrix $H_f(x_0)$ positiv definit ist, hat f im Punkt x_0 ein Minimum.
b) Falls $H_f(x_0)$ negativ definit ist, hat f im Punkt x_0 ein Maximum.
c) Falls $H_f(x_0)$ indefinit ist, hat f im Punkt x_0 einen Sattelpunkt.

■

Damit uns das etwas nützt, brauchen wir jetzt nur noch ein Kriterium, mit dem wir feststellen können, welche Definitheitseigenschaften die HESSE-Matrix hat. Dazu erinnern wir uns daran, daß die HESSE-Matrix für Funktionen aus $\mathcal{C}^2(D, \mathbb{R})$ nach dem SCHWARZschen Lemma symmetrisch ist, und wie wir aus der Linearen Algebra wissen, ist jede symmetrische Matrix diagonalisierbar.

Für eine Diagonalmatrix A mit Einträgen $\lambda_1, \ldots, \lambda_n$ und einen Vektor v mit Komponenten $v_1, \ldots, v_n$ wird obige quadratische Form zu

$$(v_1, v_2, \ldots, v_n) \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_n \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} = \lambda_1 v_1^2 + \ldots + \lambda_n v_n^2 \, ;$$

eine Diagonalmatrix ist also genau dann positiv definit, wenn alle Diagonaleinträge positiv sind und genau dann negativ definit, wenn sie alle negativ sind. Falls es sowohl positive als auch negative Diagonaleinträge gibt, ist die Matrix indefinit.

Nun ist es für den Wertebereich einer Funktion irrelevant, bezüglich welches Koordinatensystems wir die Argumente ausdrücken; wir können eine symmetrische Matrix also bezüglich einer Basis aus Eigenvektoren betrachten, wo sie zur Diagonalmatrix wird mit den Eigenwerten als Einträgen. Daher gilt:

Lemma: Eine symmetrische Matrix ist genau dann positiv definit, wenn alle ihre Eigenwerte positiv sind und genau dann negativ definit, wenn alle ihre Eigenwerte negativ sind. Falls es sowohl positive als auch negative Eigenwerte gibt, ist sie indefinit.

■

Da die Determinante einer Matrix gleich dem Produkt ihrer Eigenwerte

ist, folgt, daß eine Matrix nur dann positiv definit sein kann, wenn ihre Determinante positiv ist; für negativ definite $n \times n$-Matrizen muß die Determinante bei geradem n ebenfalls positiv sein, bei ungeradem negativ.

Für symmetrische 2×2-Matrizen läßt sich daraus leicht ein notwendiges und hinreichendes Kriterium machen: Das charakteristische Polynom von

$$A = \begin{pmatrix} a & b \\ b & d \end{pmatrix}$$

mit Eigenwerten λ_1 und λ_2 ist

$$\lambda^2 - (a + d)\lambda + (ad - b^2) = (\lambda - \lambda_1)(\lambda - \lambda_2)\,;$$

daher ist $\lambda_1 + \lambda_2 = a + d$.

(In der Tat rechnet man auf genau die gleiche Weise leicht nach, daß für jede $n \times n$-Matrix die Summe der n Eigenwerte gleich der Summe der n Diagonaleinträge ist, die sogenannte *Spur* der Matrix.)

Wenn $\det A = ad - b^2$ positiv ist, haben nicht nur λ_1 und λ_2, sondern auch a und d dasselbe Vorzeichen, das somit gleich dem von $a + d = \lambda_1 + \lambda_2$ ist. Als Zusammenfassung der obigen Diskussion können wir daher festhalten

Satz: Eine symmetrische reelle 2×2-Matrix A ist genau dann positiv definit, wenn $\det A > 0$ und $a > 0$ ist, negativ definit, wenn $\det A > 0$ und $a < 0$ ist, und indefinit wenn $\det A < 0$ ist. ∎

(Anstelle von a könnte hier natürlich überall auch d stehen.)

Beispielsweise ist die Matrix $\begin{pmatrix} 1 & 2 \\ 2 & 5 \end{pmatrix}$ positiv definit, denn sie hat Determinante eins und positive Diagonaleinträge. Im obigen Beispiel des Sattelpunkts mit $f(x, y) = x^2 - y^2$ ist

$$H_f(0, 0) = \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix}$$

offensichtlich indefinit, was man nicht nur an der negativen Determinanten sieht.

d) Lineare Regression

Als etwas umfangreicheres Beispiel für die Anwendung des obigen Satzes wollen wir ein klassisches Problem aus der Statistik betrachten, die Suche nach einer sogenannten *Ausgleichskurve* durch eine gegebene Punktmenge. Einige werden vielleicht aus dem Physikunterricht mit Ausgleichsgeraden vertraut sein: Wenn zwei physikalische Größen in einem linearen Zusammenhang stehen, werden trotzdem die zugehörigen Meßgrößen auf Grund die lineare Gleichung im allgemeinen nicht erfüllen: Messungen sind praktisch immer mit Fehlern behaftet. Bei wirtschafts- und sozialwissenschaftlichen Daten sind exakte Gesetze ohnehin die Ausnahme; hier kann man nur auf näherungsweise Gültigkeit hoffen.

Das Problem, zu vorgegebenen Daten einen möglichst guten näherungsweisen Zusammenhang zu finden, bezeichnet man als *Regression;* falls sich die Koeffizienten der gesuchten Funktion durch Lösen eines linearen Gleichungssystems aus den Daten berechnen lassen, redet man von *linearer Regression.* (Der Zusammenhang zwischen den Daten muß also *nicht* linear sein.)

Ausgangspunkt für die Lösung solcher Probleme ist die näherungsweise Lösung von linearen Gleichungssystemen: Wenn wir ein System aus N Gleichungen in m Unbekannten

$$\sum_{j=1}^{m} a_{ij} x_j = b_i \,, \qquad i = 1, \ldots, N$$

haben mit N wesentlich größer als m, können wir nur in seltenen Ausnahmefällen erwarten, daß es eine Lösung gibt. Wenn wir allerdings davon ausgehen, daß unsere Daten x_i ohnehin fehlerbehaftet sind, sollten auch gar nicht erst nach einer exakten Lösung suchen, sondern uns begnügen mit einem Tupel $(x_1, \ldots, x_m)$, für das die Gleichungen „einigermaßen“ erfüllt sind. Diese schwammige Formulierung läßt sich auf viele, nicht äquivalente Weisen präzisieren; die einfachste geht auf CARL FRIEDRICH GAUSS zurück: Wenn wir auf den linken Seiten aller Gleichungen Werte für die x_j einsetzen, erhalten wir einen Vektor aus $\mathbb{R}^N$; genauso bilden auch die rechten Seiten b_i einen Vektor aus $\mathbb{R}^N$. Falls das Gleichungssystem lösbar ist und wir eine Lösung

einsetzen, stimmen die beiden Vektoren überein. Wenn es keine exakte Lösung gibt, können wir stattdessen fordern, daß die (EUKLIDische) Länge des Differenzvektors möglichst klein sein soll. Das ist äquivalent zur Forderung, daß das Quadrat der Länge möglichst klein sein soll, d.h.

$$\sum_{i=1}^{N} \left(\sum_{j=1}^{m} a_{ij} x_j - b_i \right)^2$$

soll im Punkt $(x_1, \ldots, x_m)$ ein Minimum annehmen. Da es sich hier um eine Summe von Quadraten handelt, wird dieser Ansatz auch als *Methode der kleinsten Quadrate* bezeichnet.

Als quadratische Funktion in den x_j ist die obige Summe natürlich beliebig oft differenzierbar; nach dem Satz aus dem vorigen Abschnitt muß also im Minimum der Gradient verschwinden. Die partielle Ableitung nach x_k ist

$$\sum_{i=1}^{N} 2a_{ik} \left(\sum_{j=1}^{m} a_{ij} x_j - b_i \right) = 2 \sum_{j=1}^{m} \sum_{i=1}^{N} a_{ik} a_{ij} x_j - 2 \sum_{i=1}^{N} a_{ik} b_i \,.$$

Falls es ein Minimum gibt, muß dieses also eine Lösung des linearen Gleichungssystems

$$\sum_{j=1}^{m} \left(\sum_{i=1}^{N} a_{ik} a_{ij} \right) x_j = \sum_{i=1}^{N} a_{ik} b_i \qquad \text{für alle } k = 1, \ldots, m$$

aus m Gleichungen in m Unbekannten sein.

Schreiben wir das Ausgangssystem in Matrixform als $Ax = b$ mit

$$A = \begin{pmatrix} a_{11} & \ldots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{N1} & \ldots & a_{Nm} \end{pmatrix}, \quad b = \begin{pmatrix} b_1 \\ \vdots \\ b_N \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix},$$

so läßt sich das neue Gleichungssystem schreiben als

$$(A^T A)x = A^T b \,,$$

wir müssen also das ursprüngliche System einfach mit der transponierten

Matrix

$$A^T = \begin{pmatrix} a_{11} & \dots & a_{N1} \\ \vdots & \ddots & \vdots \\ a_{1m} & \dots & a_{Nm} \end{pmatrix}$$

multiplizieren.

Von der Problemstellung her ist klar, daß ein lokales Extremum einer Summe von Quadraten nur ein Minimum sein kann. Trotzdem wollen wir zur Vorsicht noch das Kriterium aus dem vorigen Abschnitt überprüfen, also die HESSE-Matrix auf Definitheit untersuchen.

Leiten wir die partielle Ableitung der Quadratsumme nach x_k weiter ab nach x_ℓ, erhalten wir

$$2\sum_{i=1}^{N} a_{ik}a_{i\ell}\,,$$

die HESSE-Matrix ist also gerade das Doppelte der Matrix A^TA. Für Definitheitseigenschaften ist der Faktor zwei natürlich ohne Bedeutung; wir können also die quadratische Form zu A^TA betrachten. Für $x \in \mathbb{R}^m$ ist

$$x^T(A^TA)x = (x^TA^T)(Ax) = (Ax)^T(Ax)$$

das Skalarprodukt des Vektors $Ax \in \mathbb{R}^N$ mit sich selbst und somit nie negativ. Es verschwindet genau dann, wenn Ax der Nullvektor in $\mathbb{R}^N$ ist, also insbesondere natürlich, wenn x der Nullvektor in $\mathbb{R}^m$ ist. Ob es Vektoren $x \neq 0$ gibt mit $Ax = 0$ hängt ab von der Matrix A: Ist $M < m$, *muß* es solche Vektoren geben; die Matrix A^TA ist also nur positiv semidefinit.

Wenn es um lineare Regression geht, ist dagegen N im allgemeinen deutlich größer als m; hier wird es nur sehr selten vorkommen, daß es Vektoren $x \neq 0$ aus $\mathbb{R}^m$ gibt mit $Ax = 0$. Mit den aus der Linearen Algebra bekannten Begriffen können wir das auch exakt formulieren: Genau dann, wenn die Matrix A kleineren Rang als m hat, gibt es solche Vektoren. Andernfalls ist $Ax = 0$ nur für $x = 0$, die Matrix A^TA ist also positiv definit. Da dann insbesondere auch ihre Determinante nicht verschwindet, folgt:

Satz: $Ax = b$ sei ein lineares Gleichungssystem aus N Gleichungen in m Unbekannten. Falls die Matrix A den Rang m hat, gibt es genau einen Vektor $x \in \mathbb{R}^m$, für den die (EUKLIDische) Länge von $Ax - b$ minimal wird. Er ist die eindeutig bestimmte Lösung des linearen Gleichungssystems $A^T Ax = A^T b$.

■

Als erstes Beispiel wollen wir den bekanntesten Spezialfall der linearen Regression betrachten, die *Ausgleichsgerade.* Hier geht es um N Datenpaare (x_i, y_i), zwischen denen wir einen Zusammenhang der Form $y_i \approx ax_i + b$ vermuten; gesucht sind diejenigen Werte a und b für die der Differenzvektor zwischen linker und rechter Seite minimale Länge hat. Im Gegensatz zu unserer sonstiger Konvention bezeichnen hier also x_i und y_i *bekannte* Größen, während a und b gesucht sind.

Das lineare Gleichungssystem, das wir näherungsweise lösen wollen, ist daher ein Gleichungssystem mit den Unbekannten a und b; es besteht aus den N Gleichungen

$$x_i a + b = y_i \qquad \text{für } i = 1, \ldots, N\,.$$

Matrix und rechte Seite sind daher

$$A = \begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_N & 1 \end{pmatrix} \quad \text{und} \quad y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}\,.$$

Nach dem gerade bewiesenen Satz müssen wir das System $A\binom{a}{b} = y$ von links mit der transponierten Matrix A^T multiplizieren; da

$$A^T A = \begin{pmatrix} x_1 & x_2 & \ldots & x_N \\ 1 & 1 & \ldots & 1 \end{pmatrix} \begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_N & 1 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^N x_i^2 & \sum_{i=1}^N x_i \\ \sum_{i=1}^N x_i & N \end{pmatrix}$$

und

$$A^T y = \begin{pmatrix} x_1 & x_2 & \ldots & x_N \\ 1 & 1 & \ldots & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^N x_i y_i \\ \sum_{i=1}^N y_i \end{pmatrix}$$

ist, müssen wir also das lineare Gleichungssystem

$$\begin{aligned}\left(\sum_{i=1}^{N} x_i^2\right) a + \left(\sum_{i=1}^{N} x_i\right) b &= \sum_{i=1}^{N} x_i y_i \\ \left(\sum_{i=1}^{N} x_i\right) a + \qquad N b &= \sum_{i=1}^{N} y_i\end{aligned}$$

lösen. Nach obigem Satz hat es genau dann eine eindeutig bestimmte Lösung, wenn der Rang der Matrix A gleich zwei ist, wenn also der Vektor mit Komponenten x_i linear unabhängig ist vom Vektor, dessen sämtliche Komponenten Einsen sind. Das ist offenbar genau dann der Fall, wenn die x_i nicht allesamt denselben Wert haben

Falls alle x_i gleich sind, ist auch unabhängig von jeder Linearer Algebra klar, daß wir keine Aussage darüber machen können, wie sich y in Abhängigkeit von x verändert – wir haben schließlich nur einen einzigen x-Wert.

Andernfalls sind a und b durch das obige Gleichungssystem eindeutig bestimmt; durch GAUSS-Elimination oder nach der CRAMERschen Regel erhalten wir die Werte

$$a = \frac{N \sum_{i=1}^{N} x_i y_i - \left(\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} y_i\right)}{N \sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2}$$

und

$$b = \frac{\left(\sum_{i=1}^{N} x_i^2\right)\left(\sum_{i=1}^{N} y_i\right) - \left(\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} x_i y_i\right)}{N \sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2}\,.$$

In konkreten Anwendungen dürfte es freilich einfacher sein, diese Formeln zu vergessen und stattdessen das lineare Gleichungssystem direkt zu lösen.

Diejenigen, die Ausgleichsgeraden aus der Schule kennen, hatten dort wohl vor allem mit Paaren (x_i, y_i) von Meßwerten zu tun, bei denen klar

war, daß sie auf Grund eines Naturgesetzes in einem linearen Zusammenhang stehen sollten; die Abweichungen zwischen y_i und $ax_i + b$ waren ausschließlich auf Meßfehler zurückzuführen.

In den Wirtschafts- und Sozialwissenschaften sind exakte Gesetze selten; trotzdem spielt Regression auch hier eine große Rolle, wenn es darum geht, ungefähre Zusammenhänge aufzustellen.

Als Beispiel betrachten wir das Problem der Korruption. Die vom ehemaligen Weltbankdirektor für Ostafrika PETER EIGEN 1993 gegründete Organisation *Transparency International* stellt jedes Jahr einen *Corruption Perceptions Index (CPI)* auf. Er beruht auf Befragungen vor allem von Geschäftsleuten, die in den untersuchten Ländern tätig sind und Auskunft geben sollen, für wie korrupt sie die dortige Regierung und Verwaltung halten. Für jedes Land wird auf Grund entsprechender Studien der letzten drei Jahre ein Indexwert zwischen 0 und 100 berechnet. Hundert bedeutet, daß den Befragten trotz ihrer Tätigkeit dort nichts über Korruption in diesem Land bekannt ist, eine Null entsprechend daß ohne Bimbes nichts geht.

Die neueste derzeit unter www.tranparency.org verfügbare Liste ist die von 2012; sie enthält 180 Staaten mit Dänemark, Finnland und Neuseeland (jeweils 90)) an der Spitze; Schlußlicht ist Somalia mit acht. Blättert man durch die Liste, drängt sich schnell der Eindruck auf, als seien am Ende vor allem arme Staaten zu finden, an der Spitze eher die reichen.

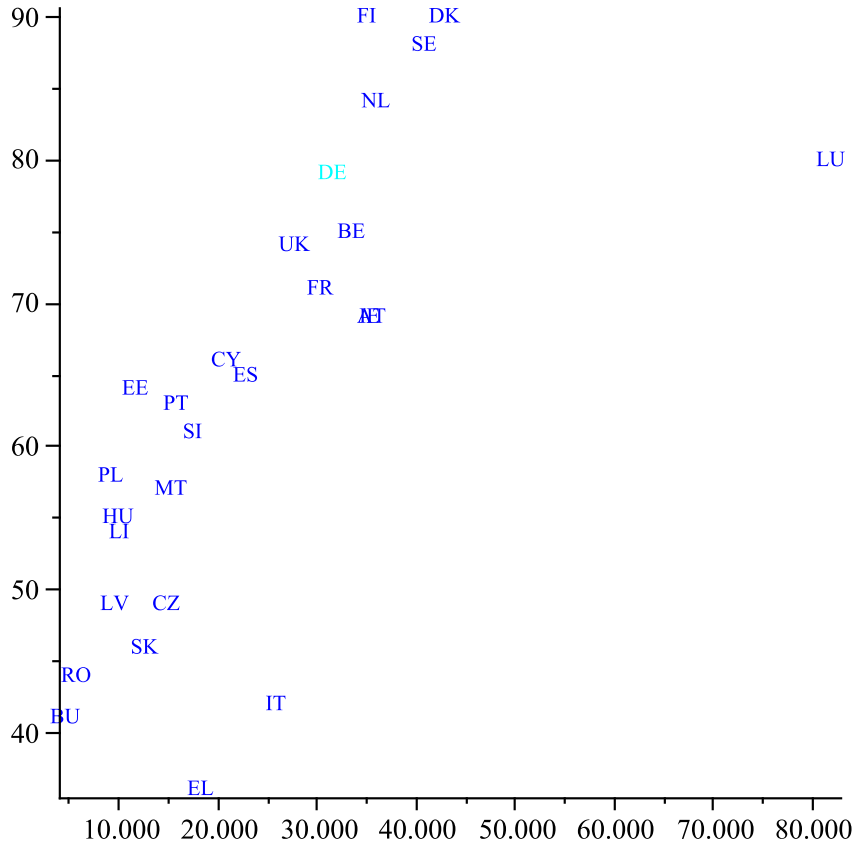
Um diese Hypothese zumindest für die 27 EU-Staaten quantitativ zu untersuchen, können wir deren CPI vergleichen mit dem Bruttoinlandsprodukt pro Einwohner.

Die Bruttoinlandsprodukte pro Einwohner (in Euro) sind bei eurostat zu finden unter ec.europa.eu/eurostat zu finden. Die derzeitig aktuellsten mehr oder weniger vollständigen Daten stammen von 2011; lediglich für Bulgarien, Polen und Rumänien liegen nur die 2010er-Werte vor. (Die kursiv gesetzten Werte für Griechenland und Portugal sind nur vorläufige Schätzungen.)

Die Tabelle auf der nächsten Seite listet für alle 27 Länder sowohl den CPI als auch das Bruttoinlandsprodukt per Einwohner auf; unten ist der Zusammenhang auch graphisch dargestellt.

Land	BIP/Einwohner	CPI
Belgien (BE)	33700	75
Bulgarien (BU)	4800	41
Dänemark (DK)	43000	90
Deutschland (DE)	31700	79
Estland (EE)	11900	64
Frankreich (FR)	30600	71
Finnland (FI)	35200	90
Griechenland (EL)	*18500*	36
Irland (IE)	35400	69
Italien (IT)	26000	42
Lettland (LV)	9800	49
Litauen (LI)	10200	54
Luxemburg (LU)	82100	80
Malta (MT)	15500	57
Niederlande (NL)	36100	84
Österreich (AT)	35700	69
Polen (PL)	9300	58
Portugal (PT)	*16000*	63
Rumänien (RO)	5800	44
Tschechische Republik (CZ)	14900	49
Schweden (SE)	41100	88
Slowakei (SK)	12700	46
Slowenien (SI)	17600	61
Spanien (ES)	23100	65
Ungarn (HU)	10000	55
Vereinigtes Königreich (UK)	27900	74
Zypern (CY)	21100	66

Wie die graphische Darstellung der Daten zeigt, liegen die Punkte ganz offensichtlich nicht auf einer Geraden, und in der Tat gibt es keinen Grund für einen festen, deterministischen Zusammenhang zwischen den beiden Größen. Trotzdem sind sie auch nicht völlig unabhängig voneinander; die dargestellte Punktwolke zeigt einen klaren Trend, der unsere Vermutung stützt, daß zumindest tendenziell in den reicheren Staaten weniger Korruption wahrgenommen wird.



Versuchen wir also, eine Ausgleichsgerade durch diese Punktwolke zu legen! Unsere x_i sind die Bruttoinlandsprodukte per Einwohner, ihre Summe ist 659 700, die Summe ihrer Quadrate 22 975 110 000. Die y_i sind die CPIs; ihre Summe ist 1 719 und die Summe der $x_i y_i$ schließlich 46 695 600. Wir bekommen also das lineare Gleichungssystem

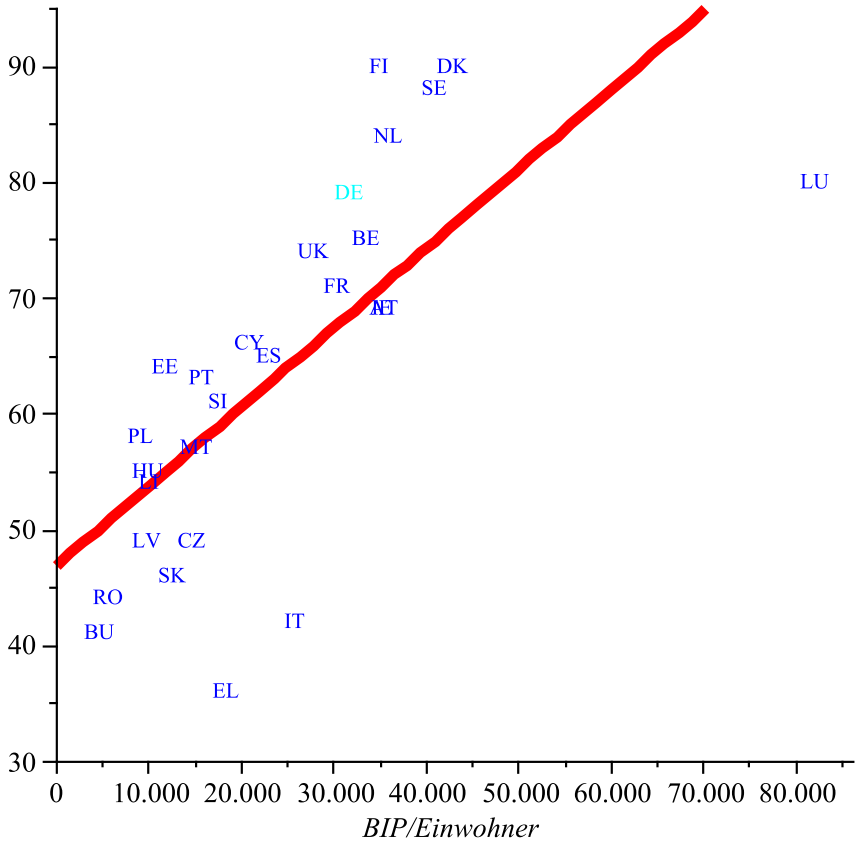
$$\begin{aligned} 22\,975\,110\,000\,a + 659\,700\,b &= 46\,695\,600 \\ 659\,700\,a + 27\,b &= 1\,719 \end{aligned}$$

mit der Lösung

$$a = \frac{15649}{22854800} \approx 6{,}847 \cdot 10^{-4} \quad \text{und} \quad b = \frac{10727317}{228548} \approx 46{,}937\,.$$

Die nächste Abbildung zeigt die Daten zusammen mit der Geraden CPI $= a \cdot$ BIP $+ b$. Sie wird zwar ihrem Namen *Ausgleichsgerade* durchaus gerecht, ist aber keine optimale Beschreibung des Zusammenhangs zwischen den beiden Größen. Insbesondere fällt ins Auge, daß die Werte für Finnland, Schweden und Dänemark ziemlich weit von der Geraden

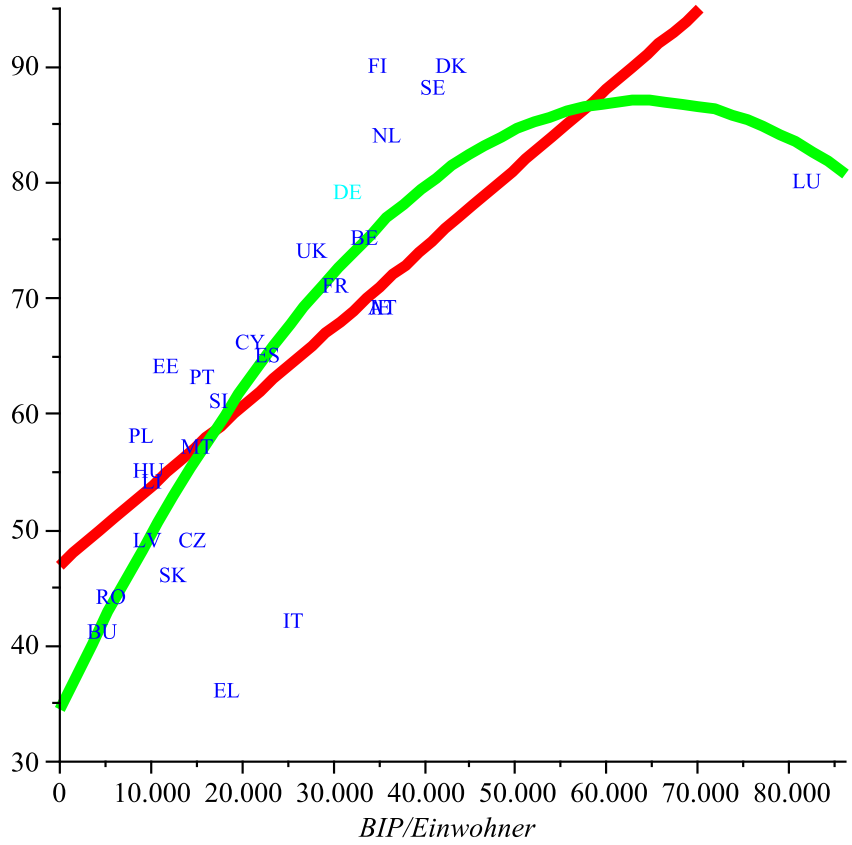
entfernt sind; noch schlimmer ist es bei Luxemburg, wo uns die Geradengleichung einen gar nicht existierenden (und auch nicht dargestellten) Wert von knapp über 103 vorhersagen würde oder bei Griechenland und Italien, die ebenfalls sehr weit unterhalb der Gerade liegen.



Nun gibt es wie gesagt keinen vernünftigen Grund, daß es hier einen linearen Zusammenhang geben sollte: Da der CPI nur Werte bis zu hundert annehmen kann, es aber keine absolute Obergrenze für das Bruttonationaleinkommen pro Einwohner gibt, kann es in der Tat schon aus theoretischen Gründen keinen solchen Zusammenhang geben. Wir sollten daher versuchen, an Stelle einer Geraden eine andere Kurve zu finden, zum Beispiel eine Parabel:

Auch die bestmögliche „Ausgleichsparabel“ läßt sich mittels linearer Regression bestimmen: Wir suchen nach einem Zusammenhang der Art

$$y_i = ax_i^2 + bx_i + c \qquad \text{für alle } i = 1, \ldots, N\,,$$



und diese N Gleichungen liefern uns ein (im allgemeinen natürlich unlösbares) lineares Gleichungssystem für die drei Koeffizienten a, b und c. Einzelheiten seien dem Leser überlassen *(siehe Übungsblatt!);* hier ist deshalb nur das Ergebnis dargestellt. Selbstverständlich ist die Übereinstimmung auch hier alles andere als perfekt, aber sie sieht doch schon besser aus als im Falle der Geraden.

Dies führt uns auf die Frage, wie wir, möglichst schon vor der Berechnung einer Ausgleichsgeraden und ohne Aufzeichnen der Datenpunkte, entscheiden können, ob es sich überhaupt lohnt. nach einem linearen Zusammenhang zu suchen.

Dazu betrachten wir als ersten Extremfall N Paare (x_i, y_i), zwischen denen ein perfekter linearer Zusammenhang besteht: $y_i = ax_i + b$ für alle i mit $a \neq 0$. Für die Mittelwerte

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \quad \text{und} \quad \overline{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

ist dann

$$\overline{y} = \frac{1}{N}\sum_{i=1}^{N} y_i = \frac{1}{N}\sum_{i=1}^{N}(ax_i + b) = a \cdot \frac{1}{N}\sum_{i=1}^{N} x_i + \frac{1}{N} \cdot Nb = a\overline{x} + b\,,$$

die Abweichung eines Werts y_i vom Mittelwert $\overline{y}$ läßt sich also mittels der Formel

$$y_i - \overline{y} = a\big(x_i - \overline{x}\big)$$

leicht aus der entsprechenden Abweichung $x_i - \overline{x}$ berechnen. Insbesondere hat das Produkt

$$\big(y_i - \overline{y}\big)\big(x_i - \overline{x}\big) = a\big(x_i - \overline{x}\big)^2$$

für alle i mit $x_i \neq \overline{x}$ dasselbe Vorzeichen, und zwar das Vorzeichen von a. Die Summe

$$\sum_{i=1}^{N}\big(y_i - \overline{y}\big)\big(x_i - \overline{x}\big) = a\sum_{i=1}^{N}\big(x_i - \overline{x}\big)^2 = \frac{1}{a}\sum_{i=1}^{N}\big(y_i - \overline{y}\big)^2\,.$$

Somit ist

$$\left(\sum_{i=1}^{N}\big(y_i - \overline{y}\big)\big(x_i - \overline{x}\big)\right)^2 = \left(\sum_{i=1}^{N}\big(y_i - \overline{y}\big)^2\right)\left(\sum_{i=1}^{N}\big(x_i - \overline{x}\big)^2\right)$$

oder, anders ausgedrückt,

$$\frac{\displaystyle\sum_{i=1}^{N}\big(y_i - \overline{y}\big)\big(x_i - \overline{x}\big)}{\displaystyle\sqrt{\sum_{i=1}^{N}\big(x_i - \overline{x}\big)^2}\sqrt{\sum_{i=1}^{N}\big(y_i - \overline{y}\big)^2}} = \begin{cases} 1 & \text{falls } a > 0 \\ -1 & \text{falls } a < 0 \end{cases}.$$

Als zweiten Extremfall betrachten wir Punktepaare (x_i, y_i), bei denen *keinerlei* Zusammenhang zwischen x_i und y_i besteht. Dann sollte man, zumindest bei hinreichend vielen Wertepaaren, erwarten, daß auf eine positive Differenz $x_i - \overline{x}$ ungefähr gleich oft eine positive wie eine negative Differenz $y_i - \overline{y}$ trifft und daß diese Differenzen im Mittel auch etwa denselben Betrag haben. Somit sollte

$$\sum_{i=1}^{N}\big(y_i - \overline{y}\big)\big(x_i - \overline{x}\big)$$

im Vergleich zum ersten Fall einen ziemlich kleinen Betrag haben; insbesondere sollte der Betrag des Quotienten

$$\frac{\sum_{i=1}^{N}\big(y_i-\overline{y}\big)\big(x_i-\overline{x}\big)}{\sqrt{\sum_{i=1}^{N}\big(x_i-\overline{x}\big)^2}\sqrt{\sum_{i=1}^{N}\big(y_i-\overline{y}\big)^2}}$$

deutlich kleiner sein als eins.

Definition: Für N Wertepaare (x_i, y_i), bei denen weder alle x_i noch alle y_i denselben Wert haben, bezeichnen wir

$$\kappa \underset{\text{def}}{=} \frac{\sum_{i=1}^{N}\big(y_i-\overline{y}\big)\big(x_i-\overline{x}\big)}{\sqrt{\sum_{i=1}^{N}\big(x_i-\overline{x}\big)^2}\sqrt{\sum_{i=1}^{N}\big(y_i-\overline{y}\big)^2}}$$

als den *Korrelationskoeffizienten.*

Dieser Korrelationskoeffizient liegt immer zwischen -1 und 1, denn bezeichnen wir mit $u \in \mathbb{R}^N$ den Vektor mit Komponenten $x_i - \overline{x}$ und mit $v \in \mathbb{R}^N$ den mit Komponenten $y_i - \overline{y}$, so ist offensichtlich

$$\kappa = \frac{\langle u, v\rangle}{\sqrt{\langle u, u\rangle}\sqrt{\langle v, v\rangle}} = \frac{\langle u, v\rangle}{\|u\| \cdot \|v\|} ,$$

wobei $\|\cdot\|$ die EUKLIDische Norm bezeichnet, und nach der CAUCHY-SCHWARZschen Ungleichung ist

$$|\langle u, v\rangle| \leq \|u\| \cdot \|v\|$$

mit Gleichheit genau dann, wenn u und v linear abhängig sind, wenn also zwischen den x_i und den y_i ein linearer Zusammenhang besteht.

Somit ist $\kappa = 1$ genau dann, wenn wir eine Gleichung $y_i = ax_i + b$ mit $a > 0$ haben, $\kappa = -1$ bei einer entsprechenden Gleichung mit $a < 0$, und $\kappa \approx 0$, wenn es keinerlei Zusammenhang zwischen den x_i und den y_i gibt.

Im obigen Beispiel ist $\kappa \approx 0{,}7099$, wir sind also weit von einem perfekten linearen Zusammenhang entfernt, noch weiter allerdings von zwei

Größen, zwischen denen keinerlei Zusammenhang besteht. Wenn wir die drei besonders weit von der Ausgleichsgerade wegliegenden Länder Griechenland, Italien und Luxemburg nicht berücksichtigen, sollten wir einen deutlich besseren Korrelationskoeffizienten erhalten, und in der Tat ist dann $\kappa \approx 0{,}9153$, In diesem Fall ist auch die Ausgleichsparabel fast ununterscheidbar von der Ausgleichsgeraden.

Da das Skalarprodukt zweier Vektoren gleich dem Produkt der Längen mal dem Kosinus des eingeschlossenen Winkels ist, können wir κ auch geometrisch interpretieren als den Kosinus des Winkels zwischen den Vektoren u und v. Im Falle $\kappa = \pm 1$ ist dieser Winkel gleich 0° oder 180°, die beiden Vektoren liegen also auf einer Geraden; ist $\kappa = 0$, bilden sie einen rechten Winkel. Im obigen Fall ist der Winkel ungefähr $44{,}77^\circ$ bei Berücksichtigung aller 27 Staaten und ungefähr $23{,}75^\circ$ wenn wir Griechenland, Italien und Luxemburg ausschließen.

e) Höhere Ableitungen impliziter Funktionen

Die Kriterien aus Abschnitt *c)* helfen uns bei der Klassifikation von Extrema ohne Nebenbedingungen; den Fall von Extrema unter Nebenbedingungen haben wir zumindest theoretisch via den Satz über implizite Funktionen darauf zurückgeführt.

In den meisten praktischen Anwendungen wird man bei der Lösung einer Optimierungsaufgabe unter Nebenbedingungen *ad hoc* entscheiden können, wo die Maxima und/oder Minima liegen; teilweise helfen auch die im nächsten Paragraphen behandelten Existenzsätze.

Im Prinzip können wir allerdings auch für eine implizit gegebene Funktion höhere Ableitungen berechnen – vorausgesetzt, natürlich, daß sie existieren.

Beim Satz über implizite Funktionen hatten wir eine differenzierbare Funktion $F(x, y)$ zweier Veränderlicher und einen Punkt $(x_0, y_0) \in \mathbb{R}^2$ mit $F(x_0, y_0) = 0$, aber $F_y(x_0, y_0) \neq 0$. Unter diesen Voraussetzungen konnten wir zeigen, daß es um x_0 eine differenzierbare Funktion $f(x)$ gibt, so daß $F\big(x, f(x)\big)$ identisch verschwindet. Im Beweis berechneten wir auch gleich die Ableitung $f'(x)$; sobald wir allerdings wissen, daß diese existiert, können wir sie auch einfacher bestimmen: Die Funktion $x \mapsto F\big(x, f(x)\big)$ ist gleich der Nullfunktion, und damit verschwindet

natürlich ihre Ableitung. Andererseits ist diese Ableitung nach der Kettenregel gleich

$$F_x\big(x, f(x)\big) + F_y\big(x, f(x)\big) \cdot f'(x)\,,$$

also folgt

$$f'(x) = -\frac{F_x\big(x, f(x)\big)}{F_y\big(x, f(x)\big)}\,.$$

Genauso kann nun auch die zweite Ableitung von f berechnet werden – falls sie existiert. Wenn F und f zweimal stetig differenzierbar sind, können wir $F\big(x, f(x)\big)$ zweimal ableiten, was natürlich immer noch Null ist. Nach der Kettenregel ist aber die zweite Ableitung von $F\big(x, f(x)\big)$ (der Übersichtlichkeit halber jeweils ohne das Argument $\big(x, f(x)\big)$ geschrieben) gleich

$$\begin{aligned}&\frac{\partial}{\partial x}\big(F_x + F_y \cdot f'(x)\big) + \frac{\partial}{\partial y}\big(F_x + F_y \cdot f'(x)\big) \cdot f'(x)\\ &= F_{xx} + F_{yx} \cdot f'(x) + F_y \cdot f''(x) + \big(F_{xy} + F_{yy} \cdot f'(x)\big) \cdot f'(x)\,,\end{aligned}$$

d.h.

$$\begin{aligned}f''(x) &= -\frac{1}{F_y}\big(F_{xx} + 2F_{xy} \cdot f'(x) + F_{yy} \cdot f'(x)^2\big)\\ &= -\frac{1}{F_y^3}\big(F_{xx}^2 F_y^2 - 2F_{xy}F_xF_y + F_{yy}F_x^2\big)\,.\end{aligned}$$

Für Extremwertbetrachtungen bei implizit definierten Funktionen braucht man, die zweite Ableitung vor allem in den Punkten, in denen die erste verschwindet; dort vereinfacht sich die Formel zu

$$f''(x) = -\frac{F_{xx}\big(x, f(x)\big)}{F_y\big(x, f(x)\big)} \quad \text{falls} \quad f'(x) = 0\,.$$

Entsprechend lassen sich auch zunehmend komplizierter werdende Formeln für höhere Ableitungen herleiten, und wenn F von mehr als zwei Variablen abhängt auch solche für partielle Ableitungen.

§4: Die Topologie des $\mathbb{R}^n$

Für Funktionen einer Veränderlichen haben wir Sätze wie den Zwischenwertsatz, der uns im wesentlichen sagt, daß jede stetige Funktion

ein abgeschlossenes Intervall wieder auf ein abgeschlossenes Intervall abbildet, und auch den Satz vom Maximum, wonach eine stetige Funktion auf einem abgeschlossenen Intervall sowohl ihr Infimum als auch ihr Supremum annimmt. Funktionen mehrerer Veränderlicher haben kompliziertere Definitionsbereiche als Intervalle; wir brauchen daher etwas mehr Aufwand, um auch hier analoge Sätze zu formulieren.

a) Kompakte Mengen

Kompakte Teilmengen des $\mathbb{R}^n$ sollen im wesentlichen die Rolle spielen, die abgeschlossene Intervalle in $\mathbb{R}$ spielen. Wegen der teils sehr komplizierten Gestalt der Definitionsbereiche unserer Funktionen kommen wir zu ihrer Definition allerdings nur über einen auf den ersten Blick eher seltsamen Umweg.

Im Laufe sowohl der *Analysis I* als auch der *Analysis II* war immer wieder die Rede davon, daß gewisse Aussagen in einer *kleinen* Umgebung eines Punktes x gelten; wie groß diese Umgebung ist, hat uns im Einzelnen nicht weiter interessiert; wir forderten nur, daß es irgendein $\varepsilon > 0$ geben solle, so daß alle y mit $\|x - y\| < \varepsilon$ dazu gehören. Wenn wir verschiedene Punkte x betrachten, werden wir dabei eventuell für jeden von diesen ein anderes ε haben.

Nehmen wir etwa an, wir haben in $\mathbb{R}$ die Punkte $x_n = \frac{1}{n}$ und dazu die Umgebungen $U_n = \left(\frac{1}{2n}, \frac{3}{2n}\right)$, d.h. also alle Punkte, deren Abstand von $x_n = \frac{1}{n}$ kleiner ist als $\frac{1}{2n}$. Innerhalb jeder der Mengen U_n soll irgendeine für uns interessante Aussage gelten; beispielsweise soll es möglich sein, eine Funktion $f\colon \mathbb{R} \to \mathbb{R}$ in jeder der Mengen U_n durch eine deutlich einfachere Funktion $f_n\colon U_n \to \mathbb{R}$ anzunähern, so daß der Fehler eine gewisse Schranke nicht überschreitet.

Um die Funktion auf dem offenen Intervall $(0,\, 1)$ näherungsweise zu berechnen, können wir uns für jeden Punkt $x \in (0,\, 1)$ eine Menge U_n wählen, die x enthält: Für $x \in (0,\, 1)$ ist $1/x > 1$, also liegt zwischen $1/x$ und $3/x$ mindestens eine gerade Zahl $2n$, und für diese ist

$$\frac{1}{x} < 2n < \frac{3}{x}\,, \qquad \text{also} \qquad \frac{1}{2n} < x < \frac{3}{2n}\,.$$

Wir können also immer mindestens ein n finden, für das x in U_n liegt, und $f(x)$ dann durch $f_n(x)$ annähern. Für praktische Zwecke, zum Bei-

spiel für ein Computerprogramm, ist das vor allem dann nützlich, wenn wir mit endlich vielen Mengen U_n und damit auch mit endlich vielen Näherungsfunktionen f_n auskommen können.

Das ist hier aber leider nicht möglich: In U_n liegen nur Zahlen, die größer sind als $1/2n$. Wenn wir eine endliche Auswahl $U_{n_1}, \ldots, U_{n_k}$ dieser Mengen betrachten mit $n_1 < \cdots < n_k$, so enthält also keine dieser Mengen eine Zahl kleiner oder gleich $1/2n_k$.

Hätten wir allerdings zusätzlich zu den Mengen U_n noch irgendeine offene Menge U_0, die die Null enthält, so würden endlich viele Mengen ausreichen: Da die Null ein innerer Punkt von U_0 sein müßte, gäbe es ein $\delta > 0$, so daß U_0 das Intervall $(-\delta,\ \delta)$ enthielte, und damit würde es ausreichen, nur das Intervall U_0 sowie die Intervalle U_n mit $n < 1/(2\delta)$ zu betrachten (oder sogar nur eine Auswahl davon).

Bei der praktischen Approximation von Funktionen dürfte dieses Beispiel zwar kaum eine Rolle spielen, aber es gibt eine Vielzahl von Situationen sowohl in der Analysis als auch der Geometrie und anderen Gebieten, in denen es von entscheidender Bedeutung ist, daß wir mit endlich vielen der vorgegebenen Mengen überdecken können. Deshalb ist die folgende Definition, so künstlich sich bei der ersten Lektüre auch erscheinen mag, in weiten Teilen der Mathematik von fundamentaler Bedeutung:

Definition: *a)* Ein System $\mathfrak{U} = \{U_i \mid i \in I\}$ von offenen Teilmengen $U_i \in \mathbb{R}^n$, wobei I eine beliebige Indexmenge bezeichnet, heißt *offene Überdeckung* der Teilmenge $X \subseteq \mathbb{R}^n$, wenn

$$X \subseteq \bigcup_{i \in I} U_i$$

in der Vereinigungsmenge aller U_i liegt.
b) Ist $J \subseteq I$ eine Teilmenge von I und liegt X bereits in der Vereinigung aller U_i mit $i \in J$, bezeichnen wir $\mathfrak{V} = \{U_i \mid i \in J\}$ als eine *Teilüberdeckung* von $\mathfrak{U}$. Ist speziell J eine endliche Menge, so sprechen wir von einer *endlichen Teilüberdeckung.*
c) Eine Teilmenge $X \subseteq \mathbb{R}^n$ heißt *kompakt,* wenn jede offene Überdekkung $\mathfrak{U}$ von X eine endliche Teilüberdeckung hat.

Beginnen wir zur Veranschaulichung mit einigen Beispielen von Überdeckungen; $\|\cdot\|$ soll dabei stets die EUKLIDische Norm bezeichnen:

$\mathfrak{U}_1$ bestehe aus allen offenen Kreisscheiben

$$U_x = \left\{ y \in \mathbb{R}^n \;\middle|\; \|x - y\| < 1 \right\}$$

vom Radius eins um Punkte $x \in \mathbb{R}^n$; hier ist also die Menge I der gesamte $\mathbb{R}^n$. Offensichtlich ist $\mathfrak{U}_1$ eine offene Überdeckung sowohl von $\mathbb{R}^n$ als auch von jeder Teilmenge $X \subseteq \mathbb{R}^n$. Zumindest als offene Überdeckung von ganz $\mathbb{R}^n$ hat sie keine endliche Teilüberdeckung, denn sonst gäbe es ja endlich viele Punkte, so daß jeder beliebige Punkt in $\mathbb{R}^n$ von mindestens einem dieser Punkte höchstens den Abstand eins hätte. Damit ist klar, daß $\mathbb{R}^n$ nicht kompakt ist.

$\mathfrak{U}_2$ bestehe aus denselben Kreisscheiben U_x, jetzt aber nur für Punkte x mit ganzzahligen Koordinaten, d.h. $x \in J = \mathbb{Z}^n$. Für $n \leq 3$ ist dies ebenfalls eine offene Überdeckung von $\mathbb{R}^n$, denn für einen beliebigen Punkt $y \in \mathbb{R}^n$ erhalten wir einen Punkt $x \in \mathbb{Z}^n$ mit ganzzahligen Koordinaten, indem wir einfach jede Koordinate y_i von y zur nächstgelegenen ganzen Zahl runden. In jeder einzelnen Koordinate ist der Fehler höchstens gleich $1/2$, insgesamt also höchstens

$$\sqrt{\sum_{i=1}^{n} \left(\frac{1}{2}\right)^2} = \frac{\sqrt{n}}{2},$$

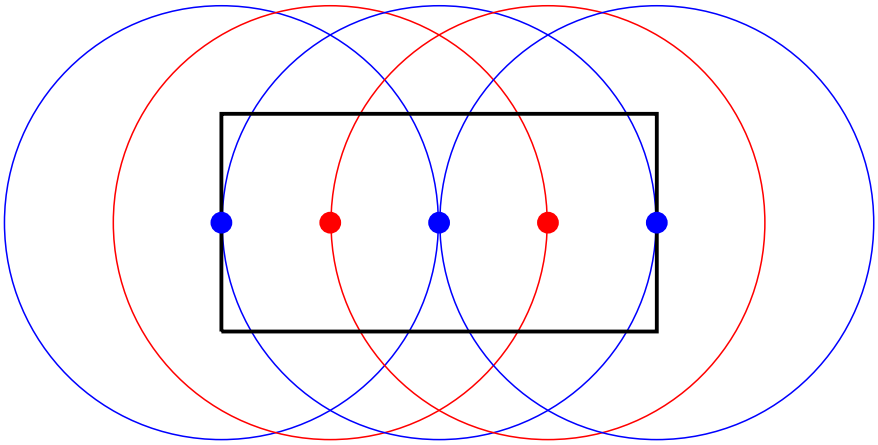
was für $n \leq 3$ kleiner als eins ist. Für $n = 4$ jedoch hat beispielsweise der Punkt $(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2})$ von jedem Punkt mit ganzzahligen Koordinaten mindestens den Abstand eins, liegt also in keiner der Mengen U_x. Somit haben wir nur für $n \leq 3$ eine Überdeckung von $\mathbb{R}^n$, die dann natürlich eine Teilüberdeckung von $\mathfrak{U}_1$ ist.

$\mathfrak{U}_3$ bestehe aus allen Intervallen der Form $\left(\frac{1}{2n}, \frac{3}{2n}\right)$ für $n \in \mathbb{N}$; hier ist also die Indexmenge $I = \mathbb{N}$. Wie wir oben gesehen haben, ist $\mathfrak{U}_3$ eine offene Überdeckung des offenen Intervalls $(0, 1)$, die keine endliche Teilüberdeckung hat; somit ist das offene Intervall $(0, 1)$ nicht kompakt.

$\mathfrak{U}_4$ schließlich bestehe aus den offenen Kreisscheiben vom Radius zwei um die Punkte $(i, 1)$ mit $i = 0, 1, 2, 3, 4$. Dies ist eine offene Überdeckung des Rechtecks

$$X = [0, 4] \times [0, 2] = \left\{ (x, y) \in \mathbb{R}^2 \;\middle|\; 0 \leq x \leq 4 \text{ und } 0 \leq y \leq 2 \right\}.$$

Teilüberdeckungen sind zum Beispiel die Überdeckung, aus den beiden (roten) Kreisscheiben um die Punkte $(1, 1)$ und $(3, 1)$, aber auch die Überdeckung aus den drei (blauen) Kreisscheiben um $(0, 1)$, $(2, 1)$ und $(4, 1)$.



Die obigen Beispiele haben uns gezeigt, daß $\mathbb{R}^n$ und das offene Intervall $(0,\ 1)$ *nicht* kompakt sind; für Beispiele kompakter Mengen reicht es natürlich nicht aus, nur spezielle Überdeckungen zu betrachten; hier müssen wir zeigen, daß *jede* irgendwie gegebene Überdeckung eine endliche Teilüberdeckung hat.

Wie eingangs erwähnt, sollen kompakte Mengen im $\mathbb{R}^n$ ähnliche Eigenschaften haben wie abgeschlossene Intervalle in $\mathbb{R}$; wenn dies mit obiger Definition der Fall ist, sollten insbesondere alle abgeschlossenen Intervalle kompakt sind. Wir beweisen gleich etwas mehr:

Lemma: Jeder Quader

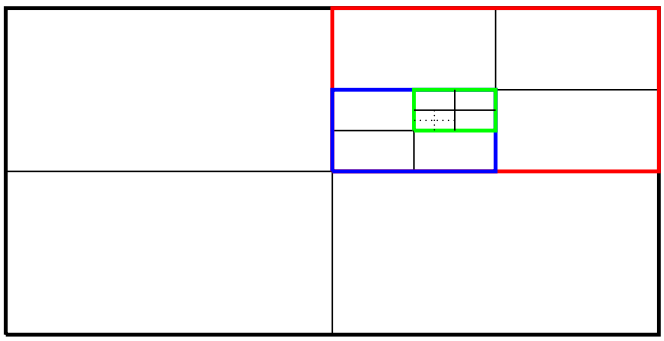
$$\begin{aligned} Q &= [a_1,\ b_1] \times \cdots \times [a_n,\ b_n] \\ &= \left\{(x_1, \ldots, x_n) \in \mathbb{R}^n \;\middle|\; a_i \leq x_i \leq b_i \ \text{ für alle } i = 1, \ldots, n\right\} \end{aligned}$$

in $\mathbb{R}^n$ ist kompakt.

Beweis: Wir nehmen an, es gebe eine Überdeckung $\mathfrak{U} = \{U_i \mid i \in I\}$ von Q, die *keine* endliche Teilüberdeckung habe, und wollen daraus einen Widerspruch herleiten, indem wir eine Art mehrdimensionale Intervallschachtelung konstruieren. Startpunkt ist der Quader Q, den wir zu diesem Zweck als $Q^{(1)} = [a_1^{(1)},\ b_1^{(1)}] \times \cdots \times [a_n^{(1)},\ b_n^{(1)}]$ schreiben. Nach Voraussetzung kann er nicht durch endlich viele Mengen U_i aus $\mathfrak{U}$ überdeckt werden.

Um aus einem Quader $Q^{(k)}$ dessen Nachfolger $Q^{(k+1)}$ zu konstruieren, teilen wir jedes der Intervalle $[a_j^{(k)}, b_j^{(k)}]$ bei seinem Mittelpunkt Mittelpunkt $c_j^{(k)} = \frac{1}{2}(a_j^{(k)} + b_j^{(k)})$ in die beiden Halbintervalle $[a_j^{(k)}, c_j^{(k)}]$ und $[c_j^{(k)}, b_j^{(k)}]$. Damit können wir den Quader $Q^{(k)}$ in 2^n Teilquader zerlegen, die sich jeweils als Produkte von n solchen Teilintervallen darstellen lassen.

Wenn $Q^{(k)}$ nicht durch endlich viele der Mengen aus $\mathfrak{U}$ überdeckt werden kann, muß für mindestens einen dieser 2^n Teilquader dasselbe gelten: Da die Vereinigung aller Teilquader gleich $Q^{(k)}$ ist, hätten wir sonst auch eine endliche Teilüberdeckung von $Q^{(k)}$. Einen solchen Quader, für den es keine endliche Teilüberdeckung gibt, bezeichnen wir als $Q^{(k+1)}$ und zerteilen ihn weiter.



$$Q = Q^{(1)} \supset Q^{(2)} \supset Q^{(3)} \supset Q^{(4)} \supset \cdots$$

Damit haben wir eine Folge von Quadern $Q^{(1)} \supset Q^{(2)} \supset Q^{(3)} \supset \cdots$, von denen keiner durch endlich viele der U_i überdeckt werden kann. In den n Koordinaten haben wir jeweils eine entsprechende Folge von Intervallen

$$[a_j^{(1)}, b_j^{(1)}] \supset [a_j^{(2)}, b_j^{(2)}] \supset [a_j^{(3)}, b_j^{(3)}] \supset \cdots,$$

von denen jedes die halbe Länge hat wie sein Vorgänger. Damit geht die Länge $b_j^{(k)} - a_j^{(k)}$ für $k \to \infty$ gegen Null, die obige Folge von Intervallen ist also eine Intervallschachtelung und definiert somit eine reelle Zahl x_j.

Wir betrachten den Punkt $(x_1, \ldots, x_n) \in \mathbb{R}^n$. Nach Konstruktion liegt er in jedem der Quader $Q^{(k)}$, insbesondere also in Q selbst. Da Q in der Vereinigung der Mengen U_i liegt, muß es daher ein $i \in I$ geben, so daß x in U_i liegt.

Da U_i eine offene Menge ist, gibt es ein $\delta > 0$, so daß mit x auch jedes $y \in \mathbb{R}^n$ mit $\|y - x\| < \delta$ in U_i liegt. Damit müssen aber auch ab einem gewissen k_0 alle $Q^{(k)}$ mit $k \geq k_0$ in U_i liegen: Der Ausgangsquader $Q^{(1)}$ hat den Durchmesser

$$d = \sqrt{(b_1 - a_1)^2 + \cdots + (b_n - a_n)^2}\,,$$

und da alle Kanten von $Q^{(k+1)}$ genau halb so lang sind wie die entsprechenden Kanten von $Q^{(k)}$, ist auch der Durchmesser nur halb so lang, d.h. der Durchmesser von $Q^{(k)}$ ist $d/2^{k-1}$, und das ist ab einem gewissen k_0 kleiner als δ. Somit hat für $k \geq k_0$ jeder von $Q^{(k)}$ höchstens den Abstand $d/2^{k-1}$ von x, also einen kleineren Abstand als δ. Dies wiederum bedeutet, daß $Q^{(k)}$ in U_i liegt, im Widerspruch zur Annahme, daß $Q^{(k)}$ nicht durch endlich viele der Mengen aus $\mathfrak{U}$ überdeckt werden kann. Damit ist das Lemma bewiesen. ∎

Wenn Quader die einzigen kompakten Teilmengen von $\mathbb{R}^n$ wären, hätte sich der Aufwand für die Definition einen so komplizierten Begriffs nicht gelohnt. Das folgende Lemma liefert uns zusammen mit dem gerade bewiesenen eine Fülle von weiteren Beispielen, die insbesondere auch krummlinig begrenzt sein können:

Lemma: Ist die abgeschlossene Menge $Z \subset \mathbb{R}^n$ Teilmenge einer kompakten Menge $K \subset \mathbb{R}^n$, ist auch Z kompakt.

Beweis: $\mathfrak{U} = \{U_i \mid i \in I\}$ sei eine offene Überdeckung der Menge Z. Wegen der Abgeschlossenheit von Z ist $V = \mathbb{R}^n \setminus Z$ offen; nehmen wir V noch mit dazu, erhalten wir eine offene Überdeckung $\mathfrak{V} = \mathfrak{U} \cup \{V\}$ von K, denn jeder Punkt aus $K \setminus Z$ liegt erst recht in $V = \mathbb{R}^n \setminus Z$, und jeder Punkt aus Z liegt in mindestens einer der Mengen U_i.

Da K kompakt ist, hat $\mathfrak{V}$ eine endliche Teilüberdeckung (von K). Wenn diese Teilüberdeckung ohne die Menge V auskommt, ist sie gleichzeitig eine endliche Teilüberdeckung von $\mathfrak{U}$ für die Menge Z. Andernfalls betrachten wir die Teilüberdeckung *ohne* die Menge V. Das ist dann eine endliche Teilmenge von $\mathfrak{U}$, und es ist auch eine offene Überdeckung von Z, denn jeder Punkt von $Z \subseteq K$ muß in einer der offenen Mengen aus der Teilüberdeckung liegen, und er liegt sicher nicht in $V = \mathbb{R}^n \setminus Z$. Somit hat $\mathfrak{U}$ eine endliche Teilüberdeckung von Z. ∎

Damit kennen wir im wesentlichen bereits alle kompakten Teilmengen von $\mathbb{R}^n$:

Definition: Eine Teilmenge $X \subseteq \mathbb{R}^n$ heißt *beschränkt*, wenn es ein $M \in \mathbb{R}$ gibt, so daß $\|x\| \leq M$ ist für alle $x \in X$.

Da die beiden Normen äquivalent sind, ist es hierbei unwesentlich, ob wir mit der EUKLIDischen Norm oder der Maximumsnorm arbeiten.

Satz von Heine-Borel: Eine Teilmenge $X \subseteq \mathbb{R}^n$ ist genau dann kompakt, wenn sie abgeschlossen und beschränkt ist.

Beweis: Sei zunächst $X \subseteq \mathbb{R}^n$ kompakt. Wir betrachten die offene Überdeckung von X aus den Mengen

$$U_x = \left\{ y \in \mathbb{R}^n \;\middle|\; \|y - x\| < 1 \right\}.$$

Wegen der Kompaktheit von x hat sie eine endliche Teilüberdeckung; diese bestehe aus den Mengen $U_{x_1}, \ldots, U_{x_r}$. Nach der Dreiecksungleichung ist

$$\|y\| \leq \|x_i\| + \|y - x_i\| \leq \|x_i\| + 1 \quad \text{für alle } y \in U_{x_i};$$

bezeichnet R die größte unter den endlich vielen Normen $\|x_i\|$, ist also $\|y\| \leq R + 1$ für alle $y \in X$. Somit ist X beschränkt.

Um zu sehen, daß X auch abgeschlossen ist, zeigen wir, daß das Komplement $\mathbb{R}^n \setminus X$ offen ist. Dazu sei $z \in \mathbb{R}^n \setminus X$ ein beliebiger Punkt aus diesem Komplement; wir müssen zeigen, daß es ein $\varepsilon > 0$ gibt, so daß $\{y \in \mathbb{R}^n \mid \|y - z\| < \varepsilon\}$ ganz in $\mathbb{R}^n \setminus X$ liegt.

Die offenen Mengen

$$U_k = \left\{ y \in \mathbb{R}^n \;\middle|\; \|z - y\| > 1/k \right\}$$

überdecken $\mathbb{R}^n \setminus \{z\}$, denn für jeden Punkt $y \neq z$ gibt es ein $k \in \mathbb{N}$, so daß die Norm von $y - z$ größer ist als $1/k$. Damit überdecken Sie insbesondere auch X, und wegen der Kompaktheit von X reicht dazu bereits eine endliche Teilüberdeckung bestehend aus gewissen Mengen $U_{k_1}, \ldots, U_{k_r}$. Ist k_r der größte unter den r Indizes, ist die Vereinigung dieser Mengen gleich U_{k_r}, d.h. $X \subseteq U_{k_r}$. Damit hat jeder Punkt aus X von z einen größeren Abstand als $1/k_r$, wir können also $\varepsilon = 1/k_r$ setzen.

Umgekehrt sei die Teilmenge $X \subseteq \mathbb{R}^n$ abgeschlossen und beschränkt. Wegen der Beschränktheit gibt es einen Quader Q, der X enthält. Dieser Quader ist nach dem ersten der obigen Lemmata kompakt, und nach dem zweiten gilt dasselbe für jede darin enthaltene abgeschlossene Teilmenge. Somit ist X kompakt.

■

HEINRICH EDUARD HEINE (1821–1881) wurde in Berlin als achtes der neun Kinder eines Bankiers geboren. Ab 1838 studierte er zunächst an der Universität Berlin, wechselte aber schon nach zum zweiten Semester nach Göttingen, wo er unter anderem Vorlesungen von GAUSS über Zahlentheorie hörte. Drei Semester später kehrte er nach Berlin zurück, wo er 1842 promovierte. Nach einem kurzen Aufenthalt an der Universität Königsberg habilitierte er sich 1844 an der Universität Bonn, wo er zunächst als Privatdozent, dann als außerplanmäßiger Professor lehrte. 1856 bekam er einen Lehrstuhl an der Universität Halle, den er bis zu seinem Tod innehatte. Seine Arbeiten befassen sich unter anderem mit partiellen Differentialgleichungen, Kettenbrüchen und elliptischen Funktionen; auch der Begriff der gleichmäßigen Stetigkeit geht auf ihn zurück.

FÉLIX EDOUARD JUSTIN EMILE BOREL (1871–1956), kurz EMILE BOREL, wurde im französischen Saint Affrique nahe der Pyrenäen als Sohn eines protestantischen Pfarrers geboren. Mit elf Jahren verließ er Saint Affrique, um zunächst in Montauben, dann in Paris weiterführende Schulen zu besuchen. Er legte sowohl die Aufnahmeprüfung zur Ecole Polytechnique als auch die zur Ecole Normale als Bester seines Jahrgangs ab und entschied sich dann zum Studium an der Ecole Normale, wo er 1893 promovierte. Danach arbeitete er als Maître de Conférence zunächst an der Universität Lille, dann an der Ecole Normale. Nach einigen weiteren Positionen unter anderem am Collège de France erhielt er 1909 einen Lehrstuhl an der Sorbonne. Trotz vielfältiger politischer Aktivitäten unter anderen als Marineminister von 1925 bis 1940 behielt er diesen Lehrstuhl bis zu seiner Verhaftung 1941 wegen seines Kampfs in der Resistance. Nach dem Krieg war er unter anderem Präsident des Wissenschaftsrats der UNESCO. Er publizierte rund zwanzig Lehrbücher und zahlreiche Arbeiten aus so unterschiedlichen Gebieten wie der reellen und komplexen Analysis, der Differentialgleichungen, der Arithmetik, der Numerik, Maßtheorie, Wahrscheinlichkeitstheorie und Spieltheorie.

Damit haben wir einen vollständigen Überblick über die kompakten Teil-

mengen von $\mathbb{R}^n$. Nach diesem Satz wird sich sicherlich mancher Leser fragen, warum man kompakte Mengen nicht einfach als abgeschlossene und beschränkte Mengen *definiert;* das wäre auf jeden Fall einfacher, als die Definition mit endlichen Teilüberdeckungen. Tatsächlich gibt es Lehrbücher der Analysis, in denen so eine Definition zu finden ist.

In der Mathematik spielt der Begriff der Kompaktheit allerdings eine sehr große Rolle nicht nur für Teilmengen des $\mathbb{R}^n$, sondern auch für viel allgemeinere Räume. Dort ist der Satz von HEINE-BOREL im allgemeinen falsch; teilweise läßt sich sogar nicht einmal definieren, was eine beschränkte Teilmenge sein soll. Im übrigen lassen sich Überdeckungen ohnehin nicht vermeiden; für die meisten Anwendungen kompakter Mengen müssen wir mit dieser Definition arbeiten. Ein Beispiel dafür ist die für uns wichtigste Anwendung kompakter Mengen, die Existenz von Maxima und Minima; diese beruht auf dem folgenden

Lemma: $f\colon D \to \mathbb{R}^m$ sei eine stetige Abbildung auf $D \subseteq \mathbb{R}^n$, und $X \subseteq D$ sei kompakt. Dann ist auch $f(X) \subseteq \mathbb{R}^m$ kompakt.

Beweis: $\mathfrak{U} = \{U_i \mid i \in I\}$ sei eine offene Überdeckung von $f(X)$. Da f eine stetige Abbildung ist, sind dann auch die Urbilder

$$f^{-1}(U_i) = \left\{x \in D \;\middle|\; f(x) \in U_i\right\}$$

offen, und natürlich bilden sie eine Überdeckung von X. Diese Überdeckung hat wegen der Kompaktheit von X eine endliche Teilüberdeckung $\{f^{-1}(U_{i_1}), \ldots, f^{-1}(U_{i_r})$. Damit überdecken $U_{i_1}, \ldots, U_{i_r}$ die Menge $f(X)$, die vorgegebene Überdeckung hat also eine endliche Teilüberdeckung.

∎

Lemma: $f\colon D \to \mathbb{R}$ sei eine stetige Abbildung auf $D \subseteq \mathbb{R}^n$, und $X \subseteq D$ sei kompakt. Dann nimmt f sowohl ihr Maximum als auch ihr Minimum an; es gibt also Elemente x_m und x_M aus X, so daß für alle $x \in X$ gilt: $f(x_m) \leq f(x) \leq f(x_M)$.

Beweis: Nach dem vorigen Lemma ist $f(X)$ eine kompakte Teilmenge von $\mathbb{R}$, also insbesondere beschränkt Somit existieren sowohl das Infimum m als auch das Supremum M von $f(X)$. Wir müssen zeigen, daß sie in $f(X)$ liegen.

Falls einer dieser beiden Punkte nicht in der abgeschlossenen Menge $f(X)$ läge, müßte er in ihrem offenen Komplement $\mathbb{R} \setminus f(X)$ liegen und hätte damit eine ε-Umgebung, die ganz in $\mathbb{R} \setminus f(X)$ läge. Im Falle des Supremums würde dies bedeuten, daß beispielsweise auch $M - \frac{\varepsilon}{2}$ eine obere Schranke von $f(X)$ wäre, im Widerspruch zur Definition des Supremums als *kleinster* oberer Schranke; im Falle des Infimums wäre entsprechend $m + \frac{\varepsilon}{2}$ eine untere Schranke.

Daher müssen m und M in $f(X)$ liegen, es gibt also Elemente x_m und x_M in X, so daß $f(x_m) = m$ und $f(x_M) = M$ ist. ∎

Als Beispiel für die Nützlichkeit dieses Lemmas wollen wir die relativen Maxima und Minima der Funktion $f(x, y) = \cos^2 x + \cos^2 y$ unter der Nebenbedingung $x^2 + y^2 \leq 1$ bestimmen.

Der Gradient von f ist $\nabla f(x, y) = \begin{pmatrix} -2 \sin x \cos x \\ -2 \sin y \cos y \end{pmatrix}$; beide Komponenten verschwinden genau dann, wenn entweder der Sinus oder der Kosinus verschwindet, wenn also x und y halbzahlige Vielfache von π sind. Da $\frac{\pi}{2}$ größer ist als eins, kommt unter der angegebenen Nebenbedingung hierfür nur der Nullpunkt in Frage. Dort ist $f(0, 0) = 2$ in der Tat ein (absolutes) Maximum der Funktion, denn der Kosinus kann nirgends größer als eins werden.

Bleiben die Extrema auf dem Rand $g(x, y) = x^2 + y^2 - 1 = 0$. Da der Gradient $\nabla g = \binom{2x}{2y}$ von g dort nirgends verschwindet, muß es für jedes solche Extremum ein λ geben mit $\nabla f(x, y) = \lambda \nabla g(x, y)$, also konkret

$$\sin x \cos x = -\lambda x \quad \text{und} \quad \sin y \cos y = -\lambda y \,.$$

Setzen wir $x = 0$ oder $y = 0$, ist jeweils eine der beiden Gleichungen erfüllt. Wegen der Nebenbedingung muß die jeweils andere Koordinate Betrag eins haben, und λ bestimmt sich aus der jeweils anderen Gleichung. Wir haben somit vier Kandidaten $(0, 1)$, $(0, -1)$ $(1, 0)$ und $(-1, 0)$. In allen vier Punkten ist

$$f(x, y) = \cos^2 0 + \cos^2 1 = 1 + \cos^2 1 \,.$$

Falls weder x noch y verschwinden, können wir dividieren und erhalten

die deutlich unangenehmeren Gleichungen

$$-\lambda = \frac{\sin x \cos x}{x} = \frac{\sin y \cos y}{y} .$$

Um diese etwas zu vereinfachen, beachten wir, daß nach EULER gilt

$$\sin x \cos x = \frac{e^{ix} - e^{-ix}}{2i} \cdot \frac{e^{ix} + e^{-ix}}{2} = \frac{e^{2ix} - e^{-2ix}}{4i} = \frac{\sin 2x}{2} ,$$

wir haben also die etwas einfachere Gleichung

$$-\lambda = \frac{\sin 2x}{2x} = \frac{\sin 2y}{2y} .$$

Wir wollen uns überlegen, daß hier $x = \pm y$ sein muß. Dazu müssen wir die Funktion $h(t) = \sin(t)/t$ genauer untersuchen. Da $h(-t) = h(t)$ ist, genügt es zu zeigen, daß für $xs = y$ sein muß, falls $x, y \geq 0$. Wir wissen bereits, daß nach DE L'HÔPITAL der Grenzwert für $t \to 0$ gleich eins ist, und wollen uns überlegen, daß die Funktion für $0 < t < \pi$ monoton fällt. Das ist genau dann der Fall, wenn ihre Ableitung

$$h'(t) = \frac{t \cos t - \sin t}{t^2}$$

dort nirgends positiv wird. Das Vorzeichen dieser Ableitung ist das ihres Zählers. Dieser verschwindet an der Stelle $t = 0$; da seine Ableitung

$$(t \cos t - \sin t)' = \cos t - t \sin t - \cos t = -t \sin t$$

für $0 < t < \pi$ negativ ist, fällt er im Intervall $(0, \pi)$ monoton, ist dort also negativ. Somit ist h dort monoton fallend; da wir nur x- und y-Werte vom Betrag höchstens eins betrachten, ist also $h(2x) = h(2y)$ nur dann möglich, wenn $y = \pm x$ ist. Eingesetzt in die Nebenbedingung führt das auf die Gleichung

$$x = \pm\frac{\sqrt{2}}{2} \quad \text{und} \quad y = \pm\frac{\sqrt{2}}{2} ;$$

wir haben also wieder vier Kandidaten, und in allen vieren haben wir denselben Funktionswert

$$f\left(\pm\frac{\sqrt{2}}{2}, \pm\frac{\sqrt{2}}{2}\right) = 2\cos^2\frac{\sqrt{2}}{2} .$$

Wir müssen noch entscheiden, wo Maxima und wo Minima angenommen werden. Da die Kreislinie offensichtlich abgeschlossen und

beschränkt ist, ist sie kompakt, wir wissen also, daß f dort sowohl sein Maximum als auch sein Minimum annimmt. Da f und g differenzierbar sind, muß dies bei einem (oder mehreren) unserer acht Kandidaten passieren. Die Funktionswerte dort sind

$$1 + \cos^2 1 \approx 1{,}29192658 \quad \text{und} \quad 2\cos^2 \frac{\sqrt{2}}{2} \approx 1{,}15594369\,;$$

daher wird in den Punkten $(\pm 1, 0)$ und $(0, \pm 1)$ das Maximum (auf dem Rand) angenommen und in den Punkten $(\pm\frac{\sqrt{2}}{2}, \pm\frac{\sqrt{2}}{2})$ das Minimum.

Das absolute Maximum auf der gesamten Kreisscheibe ist, wie wir schon wissen, die im Nullpunkt angenommene Zwei; das absolute Minimum existiert wegen der Kompaktheit der Kreisscheibe ebenfalls und muß damit in den Punkten $(\pm\frac{\sqrt{2}}{2}, \pm\frac{\sqrt{2}}{2})$ angenommen werden.

Puristen, die ohne numerische Näherungswerte auskommen möchten, können natürlich auch ohne Taschenrechner oder Computer entscheiden, welcher der beiden Werte größer ist; allerdings muß man dazu etwas tricksen. Eine Möglichkeit wäre etwa die folgende:

Da π zwischen drei und vier liegt, ist

$$\frac{\pi}{4} < 1 < \frac{\pi}{3} \Rightarrow \cos\frac{\pi}{4} > \cos 1 > \cos\frac{\pi}{3} \Rightarrow \frac{\sqrt{2}}{2} > \cos 1 > \frac{1}{2}\,,$$

also $1\frac{1}{4} < 1 + \cos^2 1 < 1\frac{1}{2}$. (Zur Erinnerung: $\frac{\pi}{4}$ entspricht $45°$ und $\frac{\pi}{3}$ ist im Winkelmaß $60°$.)

$\frac{\sqrt{2}}{2}$ liegt in der Nähe von $\frac{\pi}{4}$, also sollte $\cos\frac{\sqrt{2}}{2}$ ungefähr bei $\cos\frac{\pi}{4} = \frac{\sqrt{2}}{2}$ liegen, d.h.

$$2\cos^2\frac{\sqrt{2}}{2} \approx 2\left(\frac{\sqrt{2}}{2}\right)^2 = 1$$

sollte kleiner sein als $1 + \cos^2 1$. Nach der TAYLOR-Reihe

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots$$

des Kosinus ist $\cos\frac{\sqrt{2}}{2} = 1 - \frac{1}{2\cdot 2!} + \frac{1}{2^2\cdot 4!} - \frac{1}{2^3\cdot 6!} + \cdots$.

Da der Betrag der Summanden monoton fallend ist, muß jede Summe aus einem negativen und dem darauffolgenden positiven Summanden negativ sein, d.h.

$$\cos\frac{\sqrt{2}}{2} < 1 - \frac{1}{4} + \frac{1}{4\cdot 24} = \frac{4\cdot 24 - 24 + 1}{4\cdot 24} = \frac{73}{96}\,.$$

Um zu sehen, daß $2\cdot\cos^2\frac{\sqrt{2}}{2} < 1 + \cos^2 1$ ist, reicht es somit, wenn wir zeigen, daß

$$\left(\frac{73}{96}\right)^2 < \frac{5}{8}$$

ist. Diese Behauptung ist äquivalent zu $8\cdot 73^2 < 5\cdot 96^2$ oder $2\cdot 73^2 < 5\cdot 48^2$ oder $10658 < 11520$, also richtig – ganz in Übereinstimmung mit den numerischen Resultaten.

Mit dem Lemma, wonach eine stetige Funktion auf einer kompakten Menge sowohl ihr Maximum als auch ihr Minimum annimmt, können wir auch eine am Ende von §2*b)* aufgestellte und seither mehrfach wiederholte Behauptung beweisen:

Lemma: Alle Normen auf $\mathbb{R}^n$ sind äquivalent.

Beweis: Es genügt zu zeigem, daß jede Norm $\|\cdot\|$ äquivalent ist zur Maximumsnorm $\|\cdot\|_\infty$. Die eine Richtung ist einfach: Sind

$$e_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad e_2 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \ldots, e_n = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}$$

die Koordinateneinheitsvektoren in $\mathbb{R}^n$, so können wir $x = (x_1, \ldots, x_n)$ schreiben als $x = x_1 e_1 + \cdots + x_n e_n$; nach der Dreiecksungleichung und der sonstigen Eigenschaften einer Norm ist

$$\|x\| = \left\| \sum_{i=1}^{n} x_i e_i \right\| \leq \sum_{i=1}^{n} \|x_i e_i\| = \sum_{i=1}^{n} |x_i| \, \|e_i\| \leq \|x\|_\infty \sum_{i=1}^{n} \|e_i\| \ ,$$

denn $\|x\|_\infty$ ist ja das Maximum der Beträge der x_i. Die Summe der Normen der Einheitsvektoren ist eine positive Konstante; damit haben wir gezeigt, daß es eine solche Konstante c gibt mit der Eigenschaft, daß $\|x\| \leq c \, \|x\|_\infty$ für alle $x \in \mathbb{R}^n$.

Für die andere Abschätzung überlegen wir uns zunächst, daß jede Norm eine stetige Funktion von $\mathbb{R}^n$ nach $\mathbb{R}$ ist. Der Begriff der Stetigkeit hängt ab von einer Norm; wie stets in bisherigen Verlauf der Vorlesung arbeiten wir mit der Maximumsnorm (oder der dazu äquivalenten EUKLIDischen). Auf $\mathbb{R}$ ist das einfach der Betrag; wir müssen also zeigen, daß es zu jedem $\varepsilon > 0$ ein $\delta > 0$ gibt, so daß für zwei Punkte $x, y \in \mathbb{R}^n$ mit $\|x - y\|_\infty < \delta$ gilt: $\big|\|x\| - \|y\|\big| < \varepsilon$.

Nach der Dreiecksungleichung ist

$$\|x\| = \|y + (x - y)\| \leq \|y\| + \|x - y\| \quad \text{und}$$
$$\|y\| = \|x + (y - x)\| \leq \|x\| + \|y - x\| \ ,$$

also sind $\|x\| - \|y\|$ und $\|y\| - \|x\|$ beide kleiner als $\|x - y\| = \|y - x\|$,

und damit ist

$$|\|x\| - \|y\|| \le \|x - y\| \le c\,\|x - y\|_\infty \ .$$

Setzen wir daher $\delta = \varepsilon/c$, so ist $|\|x\| - \|y\|| < \varepsilon$ für alle $x, y \in \mathbb{R}^n$ mit $\|x - y\|_\infty < \delta$. Damit ist die Stetigkeit der Norm $\|\cdot\|$ bewiesen.

Nun betrachten wir den Würfel

$$W = \{x \in \mathbb{R}^n \mid \|x\|_\infty = 1\} \ .$$

Er ist offensichtlich abgeschlossen und beschränkt, nach dem Satz von HEINE-BOREL also kompakt.

Als stetige Funktion nimmt die Norm auf W sowohl ein Minimum als auch ein Maximum an; es gibt daher Konstanten c_1 und c_2, so daß $c_1 \le \|x\| \le c_2$ für alle $x \in W$. Beide Konstanten sind positiv, denn $\|x\|$ verschwindet nur für $x = 0$, und dieser Punkt liegt nicht in W.

Ein beliebiges $x \neq 0$ können wir schreiben als

$$x = \|x\|_\infty \cdot \frac{x}{\|x\|_\infty} \ ,$$

wobei der zweite Faktor in W liegt. Seine Norm

$$\left\| \frac{x}{\|x\|_\infty} \right\| = \left\| \frac{1}{\|x\|_\infty} \cdot x \right\| = \frac{1}{\|x\|_\infty} \|x\| = \frac{\|x\|}{\|x\|_\infty}$$

liegt zwischen c_1 und c_2, also ist

$$c_1 \le \frac{\|x\|}{\|x\|_\infty} \le c_2 \quad \text{oder} \quad c_1 \|x\|_\infty \le \|x\| \le c_2 \|x\|_\infty \ .$$

Damit ist die Äquivalenz der beiden Normen bewiesen. ∎

Als weitere Anwendung kompakter Mengen wollen wir die gleichmäßige Stetigkeit, die wir für Funktionen einer Veränderlichen in Kapitel 4 zur Konstruktion des RIEMANN-Integrals benötigten, auch für Funktionen mehrerer Veränderlicher einführen:

Definition: Eine Abbildung $f\colon D \to \mathbb{R}^m$ auf einer Teilmenge $D \subseteq \mathbb{R}^n$ heißt *gleichmäßig stetig* auf der Teilmenge $X \subseteq D$, wenn es zu jedem $\varepsilon > 0$ ein $\delta > 0$ gibt, so daß gilt: Für alle Punkte $x, y \in X$ mit $\|x - y\| < \delta$ ist $\|f(x) - f(y)\| < \varepsilon$.

Im Eindimensionalen konnten wir zeigen, daß jede stetige Funktion auf abgeschlossenen Teilintervallen ihres Definitionsbereichs gleichmäßig stetig ist; hier gilt entsprechend

Lemma: Eine stetige Abbildung $f\colon D \to \mathbb{R}^m$ auf $D \subseteq \mathbb{R}^n$ ist auf jeder kompakten Teilmenge $K \subseteq D$ gleichmäßig stetig.

Beweis: Da f auf D stetig ist, gibt es zu jedem $\varepsilon > 0$ und zu jedem x aus D ein $\delta > 0$, so daß gilt: $\|f(y) - f(x)\| < \varepsilon$, falls $\|y - x\| < \delta$. Dieses δ hängt sowohl von ε als auch von x ab. Wir müssen zeigen, daß wir zumindest für die $x \in K$ ein gemeinsames δ finden können.

Dazu halten wir ε fest und und wählen zu jedem $x \in K$ ein $\delta_x > 0$, so daß gilt: Für $\|y - x\| < \delta_x$ ist $\|f(y) - f(x)\| < \frac{1}{2}\varepsilon$. Offensichtlich bilden die Mengen

$$U_x = \left\{ y \in \mathbb{R}^n \;\middle|\; \|y - x\| < \tfrac{1}{2}\delta_x \right\}$$

eine offene Überdeckung von K; da K kompakt ist, gibt es eine Teilüberdeckung durch endlich viele Mengen $U_{x_1}, \ldots, U_{x_r}$. Wir bezeichnen die kleinste unter den r Zahlen $\frac{1}{2}\delta_{x_j}$ mit δ.

Damit liegt jedes $x \in K$ in mindestens einer der Mengen U_{x_j}. Für ein $y \in K$ mit $\|y - x\| < \delta$ ist

$$\left\|y - x_j\right\| \le \|y - x\| + \left\|x - x_j\right\| < \delta + \tfrac{1}{2}\delta_{x_j} \le \delta_{x_j} \,;$$

nach Definition von δ_{x_j} folgt daher

$$\left\|f(y) - f(x_j)\right\| < \frac{\varepsilon}{2} \quad \text{und} \quad \left\|f(x) - f(x_j)\right\| < \frac{\varepsilon}{2} \,,$$

das heißt $\|f(y) - f(x)\| < \varepsilon$. Dies zeigt die gleichmäßige Stetigkeit von f auf K.

■

b) Zusammenhängende Mengen

Die kompakten Mengen in letzten Abschnitt sollten eine Art Verallgemeinerung abgeschlossener Intervalle sein, und zumindest was die Existenz von Maxima und Minima stetiger Funktionen betrifft, leisten sie auch das, was wir von ihnen erwarteten.

Umgekehrt ist aber nicht jede kompakte Teilmenge von $\mathbb{R}$ ein abgeschlossenes Intervall: Die Vereinigung $[-4, -2] \cup [2, 4]$ ist sicherlich

abgeschlossen und beschränkt, also kompakt. Natürlich können wir nicht erwarten, daß für Funktionen auf einer solchen Menge der Zwischenwertsatz gilt. Um auch diesen aufs Mehrdimensionale zu verallgemeinern, brauchen wir einen neuen Begriff, der etwas mit der Intervalleigenschaft zu tun haben sollte.

Dafür gibt es mehrere Möglichkeiten: Ein Intervall enthält zu zwei Punkten x, y stets auch deren Verbindungsstrecke; diese Eigenschaft könnten wir auch im Mehrdimensionalen fordern. Etwas allgemeiner könnten wir aber statt einer geradlinigen Verbindung einfach *irgendeine* Verbindungskurve zwischen je zwei Punkten fordern. Schließlich könnten wir auch ganz auf Verbindungskurven verzichten und stattdessen eine andere Eigenschaft des obigen Gegenbeispiels ausnutzen: Die Vereinigung $[-4, -2] \cup [2, 4]$ ist enthalten in der Vereinigung der beiden offenen Intervalle $(-5, -1)$ und $(1, 5)$, und diese offenen Intervalle haben leeren Durchschnitt.

Alle drei Ansätze sind nützlich und haben deshalb eigene Namen:

Definition: *a)* Eine Teilmenge $X \subseteq \mathbb{R}^n$ heißt *konvex*, wenn für alle $x, y \in X$ auch deren Verbindungsstrecke ganz in X liegt.
b) X heißt *wegzusammenhängend*, wenn es für alle $x, y \in X$ eine ganz in X liegende Kurve gibt, die diese Punkte verbindet, d.h. eine stetige Abbildung $\gamma\colon D \to \mathbb{R}^n$ auf einer Teilmenge $D \subset \mathbb{R}$ mit $\gamma(0) = x$, $\gamma(1) = y$ und $\gamma\big([0, 1]\big) \subseteq X$.
c) X heißt *zusammenhängend*, wenn gilt: Sind $U, V \subseteq \mathbb{R}^n$ zwei offene Mengen mit leerem Durchschnitt und liegt X in der Vereinigung $U \cup V$, so liegt X ganz in einer der beiden Mengen.

Von diesen drei Forderungen ist die Konvexität die stärkste, der Zusammenhang die schwächste. Genauer gilt:

Lemma: *a)* Jede konvexe Teilmenge $X \subseteq \mathbb{R}^n$ ist auch wegzusammenhängend.
b) Jede wegzusammenhängende Teilmenge $X \subseteq \mathbb{R}^n$ ist auch zusammenhängend.

Beweis: a) ist klar: Wenn für je zwei Punkte $x, y \in X$ deren Verbindungsstrecke $\big\{(1-t)x + ty \;\big|\; 0 \le t \le 1\big\}$ in X liegt, können wir einfach

diese Strecke als Verbindungskurve nehmen, d.h. wir definieren

$$\gamma: \begin{cases} \mathbb{R} \to \mathbb{R}^n \\ t \mapsto (1-t)x + ty \end{cases} .$$

Dann ist $\gamma(0) = x$, $\gamma(1) = y$, und für alle $t \in [0,\ 1]$ liegt $\gamma(t)$ in X.

b) Die wegzusammenhängende Teilmenge $X \subseteq \mathbb{R}^n$ sei enthalten in der Vereinigung der beiden offenen Mengen $U, V \subseteq \mathbb{R}^n$ mit $U \cap V = \emptyset$. Falls X selbst die leere Menge ist, liegt X sowohl in U als auch in V, und wir sind fertig. Andernfalls gibt es mindestens einen Punkt $x \in X$. Dieser muß entweder in U oder in V liegen; indem wir gegebenefalls die Bezeichnungen vertauschen, können wir annehmen, daß x in U liegt. Wir müssen zeigen, daß dann auch alle anderen Punkte $y \in X$ in U liegen.

Nach Voraussetzung gibt es eine Kurve γ, die x und y miteinander verbindet. Wir wollen uns überlegen, daß diese ganz in U liegen muß. Dazu betrachten wir

$$s = \sup\left\{ u \in [0,1] \;\middle|\; \gamma(t) \in U \text{ für alle } t \in [0,u) \right\} .$$

Falls $s < 1$ wäre, könnte $\gamma(s)$ nicht in U liegen, denn wegen der Stetigkeit von γ ist $\gamma^{-1}(U)$ eine offene Menge, enthält also zu jedem ihrer Punkte auch eine offene Umgebung. s könnte aber auch nicht in V liegen, denn auch $\gamma^{-1}(V)$ ist offen, und für alle $0 \leq t < s$ liegt t in $\gamma^{-1}(U)$, also, da $U \cap V = \emptyset$, nicht in $\gamma^{-1}(V)$. Das ist ein Widerspruch, denn $X \subseteq U \cup V$, so daß $\gamma(s)$ in einer der beiden Mengen liegen muß.

Somit ist $s = 1$ und $y = \gamma(1) \in U$, denn läge $\gamma(1)$ in V, gäbe es auch eine Umgebung der Eins, so daß $\gamma(t)$ für alle t aus dieser Umgebung in V läge. Da y ein beliebiger Punkt aus X war, liegt ganz X in U und ist daher zusammenhängend.

■

Damit folgt beispielsweise, daß $\mathbb{R}^n$ für jedes n sowohl wegzusammenhängend als auch zusammenhängend ist, denn natürlich ist $\mathbb{R}^n$ konvex: Wir können zwei beliebige Punkte aus $\mathbb{R}^n$ stets durch eine Strecke miteinander verbinden. Damit folgt beispielsweise

Lemma: Ist $X \subseteq \mathbb{R}^n$ sowohl offen als auch abgeschlossen, so ist entweder $X = \mathbb{R}^n$ oder $X = \emptyset$.

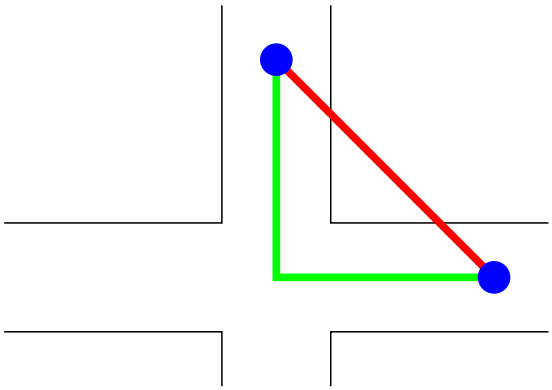
Beweis: Ist X sowohl offen als auch abgeschlossen, ist auch $\mathbb{R}^n \setminus X$ offen und hat natürlich leeren Durchschnitt mit X. Die Vereinigung dieser beiden offenen Mengen ist ganz $\mathbb{R}^n$, und da dies eine zusammenhängende Menge ist, liegt entweder ganz $\mathbb{R}^n$ in X, d.h. $X = \mathbb{R}^n$, oder aber ganz $\mathbb{R}^n$ liegt in $\mathbb{R}^n \setminus X$, was nur für $X = \emptyset$ möglich ist. ∎

Die beiden Aussagen des vorigen Lemmas lassen sich nicht umkehren, es gibt also wegzusammenhängende Mengen, die nicht konvex sind, und zusammenhängende, aber nicht wegzusammenhängende Mengen.

Als erstes Beispiel betrachten wir die Menge

$$X = \{(x, y) \in \mathbb{R}^2 \mid |x| \leq 1 \text{ oder } |y| \leq 1\}.$$

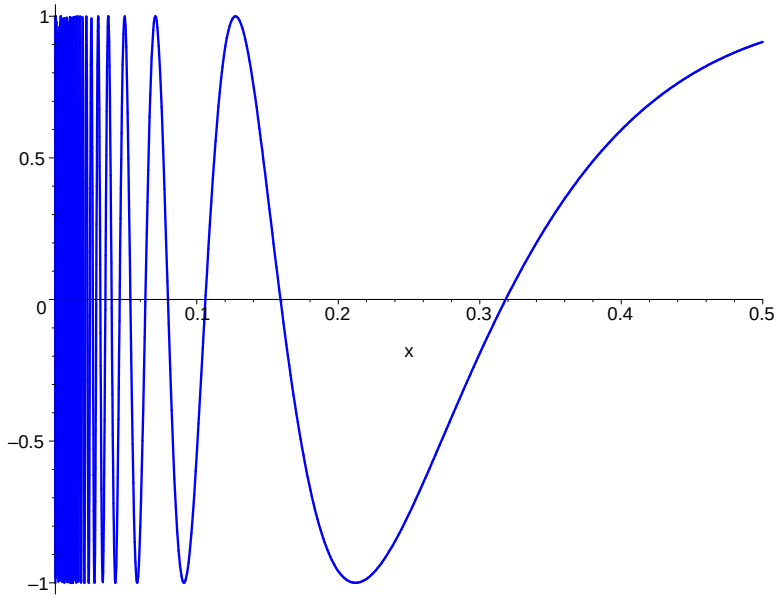
Sie ist nicht konvex, denn die Verbindungsstrecke der beiden Punkte $(0, 4)$ und $(4, 0)$ aus X enthält beispielsweise den Punkt $(2, 2)$, der nicht in X liegt. Sie ist aber wegzusammenhängend, denn für jeden Punkt $(x, y) \in X$ liegt dessen Verbindungsstrecke zum Nullpunkt in X, und für zwei Punkte aus X können wir die beiden Verbindungsstrecken zum Nullpunkt aneinandersetzen zu einer Verbindungskurve.



Beispiele zusammenhängender, aber nicht wegzusammenhängender Mengen sind schwerer zu finden; am populärsten ist die Menge

$$X = \left\{(x, \sin \tfrac{1}{x}) \in \mathbb{R}^2 \mid x > 0\right\} \cup \left\{(0, y) \mid |y| \leq 1\right\}$$

bestehend aus einer Sinuslinie mit für $x \to 0$ immer kleiner werdendem Abstand zwischen aufeinanderfolgenden Bergen und Tälern und dem Intervall auf der y-Achse, dem sich diese Sinuslinie immer mehr annähert.



Diese Menge ist nicht wegzusammenhängend; beispielsweise lassen sich die beiden Punkte $(\frac{1}{\pi}, 0)$ und $(0, 1)$ aus X nicht durch eine Kurve miteinander verbinden: Gäbe es nämlich eine Kurve γ mit $\gamma(0) = (\frac{1}{\pi}, 0)$ und $\gamma(1) = (0, 1)$, so könnten wir die Menge aller $u \in [0,\ 1]$ betrachten, für die $\gamma(t)$ im Intervall $0 \leq t < u$ überall positive x-Koordinate hat; ihr Supremum sei s. Wegen $\gamma(1) = (0, 1)$ wäre $s < 1$; außerdem müßte die x-Koordinate von $\gamma(s)$ verschwinden, denn wäre sie positiv, könnte es nicht in jeder beliebig kleinen Umgebung von $\gamma(s)$ Punkte mit x-Koordinate Null geben. Da $\sin \frac{1}{x}$ in jedem Intervall $(0,\ \varepsilon)$ alle Werte zwischen -1 und 1 annimmt, müßte $\gamma(s)$ wegen der Stetigkeit von γ in jeder beliebig kleinen Umgebung Punkte mit allen y-Koordinaten zwischen -1 und 1 haben, was natürlich nicht möglich ist. Somit ist X nicht wegzusammenhängend.

Die beiden Teilmengen

$$X_1 = \left\{(x, \sin \tfrac{1}{x}) \in \mathbb{R}^2 \;\middle|\; x > 0\right\} \quad \text{und} \quad X_2 = \left\{(0, y) \;\middle|\; |y| \leq 1\right\}$$

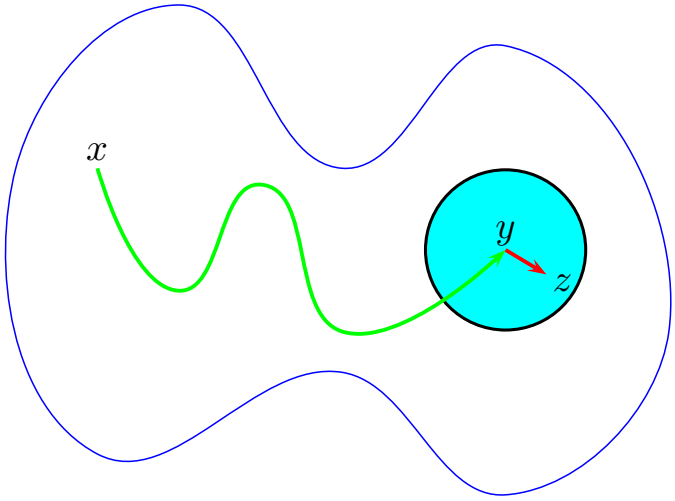
sind natürlich wegzusammenhängend und damit erst recht zusammenhängend. Wenn es zwei offene Mengen U, V mit leerem Durchschnitt gäbe, so daß $X \subseteq U \cup V$ weder in U noch in V liegt, müßte daher X_1 in U und X_2 in V liegen oder umgekehrt. Das ist aber nicht möglich, denn in jeder offenen Menge, die X_2 enthält, gibt es auch Punkte aus X_1. Daher ist X zusammenhängend.

Der große Aufwand, den wir für dieses Beispiel treiben mußten, legt

die Vermutung nahe, daß zusammenhängende Mengen in vielen Fällen auch wegzusammenhängend sind. In der Tat gilt etwa

Lemma: Eine offene Teilmenge $X \subseteq \mathbb{R}^n$ ist genau dann zusammenhängend, wenn sie wegzusammenhängend ist.

Beweis: Wir wissen bereits, daß *jede* wegzusammenhängende Menge zusammenhängend ist; zu zeigen bleibt, daß jede offene zusammenhängende Menge X auch wegzusammenhängend ist. Für die leere Menge gibt es nichts zu beweisen; sei also $X \neq \emptyset$. Wir wählen einen festen Punkt $x \in X$ und betrachten die Menge U aller $y \in X$, die durch eine Kurve mit x verbunden werden können, sowie die Menge V aller jener $y \in X$, für die das nicht der Fall ist. Beides sind offene Mengen: Wegen der Offenheit von X gibt es nämlich zu jedem $y \in X$ ein $\varepsilon > 0$, so daß auch alle $z \in \mathbb{R}^n$ mit $\|z - y\| < \varepsilon$ in X liegen. Wenn wir dabei die EUKLIDische Norm verwenden, liegen diese Punkte z in einer n-dimensionalen Kugel um y mit Radius ε und lassen sich daher alle durch eine Strecke, die ganz innerhalb der Kugel und damit auch innerhalb von X liegt, mit dem Mittelpunkt y verbinden.



Falls sich der Mittelpunkt durch eine Kurve mit x verbinden läßt, können wir diese Kurve daher um eine Strecke verlängern, um auch jeden Punkt z aus der Kugel mit x zu verbinden, so daß die gesamte Kugel in U liegt. Falls es umgekehrt innerhalb von X keine Verbindungskurve von x nach y gibt, kann es auch für kein z aus der Kugel eine solche Kurve geben, denn sonst könnten wir diese um die Verbindungsstrecke

von z nach y verlängern zu einer Verbindungskurve zwischen x und y.

Damit sind U und V offene Mengen; ihr Durchschnitt ist leer und ihre Vereinigung gleich X. Da wir X als zusammenhängend vorausgesetzt haben, muß ganz X in einer der beiden Mengen liegen, und da x in U liegt, ist das die Menge U. Dies zeigt, daß sich jeder Punkt $y \in X$ durch eine ganz in X verlaufende Kurve mit x verbinden läßt; X ist also wegzusammenhängend.

∎

Im Eindimensionalen ist die Situation noch einfacher; hier erhalten wir bei allen drei oben definierten Begriffen einfach die Intervalle, zu denen wir, wie üblich, auch die unbeschränkten Intervalle und die leere Menge rechnen:

Lemma: Für eine Teilmenge $X \subset \mathbb{R}$ sind die folgenden vier Aussagen äquivalent:
a) X ist ein Intervall.
b) X ist konvex.
c) X ist wegzusammenhängend.
d) X ist zusammenhängend.

Beweis: Um die Äquivalenz dieser vier Aussagen zu zeigen, müssen wir nicht alle $4 \cdot 3 = 12$ Implikationen einzeln nachweisen; es reicht, wenn wir in einem sogenannten *Ringschluß* die Folgerungen

$$a) \Rightarrow b) \Rightarrow c) \Rightarrow d) \Rightarrow a)$$

zeigen. Die ersten drei unter diesen sind offensichtlich *bzw.* wurden oben schon allgemein gezeigt; wirklich beweisen müssen wir daher nur, daß jede zusammenhängende Teilmenge $X \subseteq \mathbb{R}$ ein Intervall ist. Da wir die leere Menge *per definitionem* als Intervall betrachten, können wir dazu annehmen, daß X nicht leer ist.

Falls X nach oben beschränkt ist, hat X ein Supremum $b \in \mathbb{R}$; wir wollen uns als erstes überlegen, daß X dann für jedes $z \in X$ das Intervall $[z,\, b)$ enthält: Gäbe es nämlich ein $c \in [z,\, b)$, das nicht in X läge, so läge X in der Vereinigung der beiden offenen Mengen $U = \{x \in \mathbb{R} \mid x < c\}$ und $V = \{x \in \mathbb{R} \mid x > c\}$, aber weder in U noch in V. Entsprechend folgt, daß X, sofern es unbeschränkt ist, zu jedem $z \in X$ auch alle reellen

Zahlen $x \geq z$ enthalten muß, denn läge $x > z$ nicht in X, könnten wir wie eben argumentieren.

Dasselbe Argument zeigt auch, daß X, falls es nach unten beschränkt ist, für jedes $z \in X$ das Intervall $(a, z]$ enthalten muß, wobei a das Infimum von X bezeichnet; falls X nicht nach unten beschränkt ist, muß es entsprechend zu jedem $z \in X$ auch alle reellen Zahlen $x \leq z$ enthalten.

Ist also X eine beschränkte zusammenhängende Teilmenge von $\mathbb{R}$ mit Infimum a und Supremum b, so enthält X das offene Intervall (a, b). Da X keine Punkte $z < a$ oder $z > b$ enthalten kann, können dazu höchstens noch einer oder beide der Punkte a, b kommen; X ist also eines der vier Intervalle (a, b), $(a, b]$, $[a, b)$ oder $(a, b]$.

Falls X nur nach unten beschränkt ist mit Infimum a, enthält X auf jeden Fall alle reellen Zahlen $x > a$, zusätzlich eventuell nach den Punkt a. Entsprechendes gilt für eine nur nach oben beschränkte Menge.

Bleibt noch der Fall, daß X weder nach oben noch nach unten beschränkt ist; dann ist $X = \mathbb{R}$, was wir ebenfalls als Intervall betrachten. ∎

Die für uns wichtigste Anwendung zusammenhängender Mengen ist die folgende Verallgemeinerung des Zwischenwertsatzes:

Satz: Ist $f\colon D \to \mathbb{R}^m$ eine stetige Abbildung auf $D \subseteq \mathbb{R}^n$ und $X \subseteq D$ zusammenhängend, so ist auch $f(X)$ zusammenhängend.

Beweis: U und V seien zwei offene Teilmengen von $\mathbb{R}^m$ mit leerem Durchschnitt, und $f(X)$ liege in der Vereinigung $U \cup V$. Wegen der Stetigkeit von f sind die Urbilder $f^{-1}(U)$ und $f^{-1}(V)$ offene Teilmengen von $\mathbb{R}^n$; ihr Durchschnitt ist leer, denn kein $x \in D$ kann ein Bild haben, das sowohl in U als auch in V liegt. Somit muß X ganz in einer der beiden Mengen $f^{-1}(U)$ oder $f^{-1}(V)$ liegen, also $f(X)$ in U oder in V. ∎

Korollar: Ist $f\colon D \to \mathbb{R}$ eine stetige Abbildung auf $D \subseteq \mathbb{R}^n$ und $X \subseteq D$ zusammenhängend, so ist $f(X)$ ein Intervall, enthält also zu je zwei Werten a, b auch alle Zahlen, die zwischen den beiden liegen. ∎

§5: Banach-Räume

Die reellen Zahlen unterscheiden sich vor allem dadurch von den rationalen Zahlen, daß viele Folgen rationaler Zahlen, die keinen Grenzwert in $\mathbb{Q}$ haben, doch gegen einen Grenzwert aus $\mathbb{R}$ konvergieren. Insbesondere hat in $\mathbb{R}$ jede CAUCHY-Folge einen Grenzwert. Diese Eigenschaft der reellen Zahlen wollen wir in diesem Paragraphen verallgemeinern und dabei auch sehen, daß sich beispielsweise das aus Kapitel I bekannte HERON-Verfahren zur Berechnung der Quadratwurzel einordnet in eine Gruppe viel allgemeinerer Techniken.

a) Vollständigkeit

CAUCHY-Folgen und der Begriff der Vollständigkeit lassen sich für beliebige metrische Räume definieren; da wir die Begriffe nicht in dieser Allgemeinheit benötigen, beschränken wir uns auf normierte Vektorräume; die Verallgemeinerung auf metrische Räume sollte für jeden interessierten Leser offensichtlich sein.

Definition: V sei ein normierter Vektorraum mit Norm $\|\cdot\|$.
a) Eine Folge $(x_n)_{n\in\mathbb{N}}$ von Elementen $x_n \in V$ heißt CAUCHY-*Folge*, wenn es zu jedem $\varepsilon > 0$ ein $N \in \mathbb{N}$ gibt, so daß $\|x_n - x_m\| < \varepsilon$ für alle $n, m \geq N$.
b) V heißt ein *vollständiger normierter Vektorraum* oder BANACH-*Raum*, wenn jede CAUCHY-Folge aus V gegen ein Element von V konvergiert.

STEFAN BANACH (1892–1945) wurde in Krakau geboren und ausgebildet, promovierte und arbeitete dann aber an der Universität von Lvov in der Ukraine, wo er unter schwierigen Bedingungen unter deutscher Besatzung den zweiten Weltkrieg verbrachte. Durch seine Arbeiten über lineare Operatoren und über Vektorräume von Funktionen wurde er zum Begründer der modernen Funktionalanalysis. Nach dem Krieg wollte er auf einen Lehrstuhl an der Universität Krakau wechseln, starb aber 1945 an Lungenkrebs. Das wichtigste mathematische Forschungsinstitut Polens, das Banach-Zentrum in Warschau, ist nach ihm benannt.

Einfachstes Beispiel eines BANACH-Raums ist natürlich $\mathbb{R}$ selbst mit der Betragsfunktion als Norm; hier ist die Vollständigkeitsaussage gerade

das CAUCHYsche Konvergenzkriterium. Da es in diesem Semester vor allem um Funktionen mehrerer Veränderlicher geht, sollten wir uns als nächstes überlegen, ob auch $\mathbb{R}^n$ ein BANACH-Raum ist. Während wir in $\mathbb{R}$ immer mit dem Betrag arbeiten, haben wir im Mehrdimensionalen allerdings verschiedene Normen, und müssen uns, bevor wir von Vollständigkeit reden können, auf eine davon festlegen.

Zumindest für die Konvergenz von Folgen kommt es im $\mathbb{R}^n$ *nicht* darauf an, mit welcher Norm wir arbeiten: Wie wir in Abschnitt *a)* gesehen haben, sind alle Normen auf $\mathbb{R}^n$ äquivalent, und wie wir bereits aus §1*b)* wissen, führen äquivalente Normen zum gleichen Konvergenzbegriff. Dasselbe gilt auch für CAUCHY-Folgen: Sind $\|\cdot\|_1$ und $\|\cdot\|_2$ zwei äquivalente Normen auf einem $\mathbb{R}$-Vektorraum V und ist $(x_n)_{n\in\mathbb{N}}$ eine CAUCHY-Folge bezüglich $\|\cdot\|_1$, so gibt es zu jedem $\varepsilon > 0$ ein $N \in \mathbb{N}$, so daß $\|x_n - x_m\|_1 < \varepsilon$ für alle $n, m \geq N$. Wegen der Äquivalenz der beiden Normen gibt es außerdem eine positive reelle Zahl c, so daß $\|x\|_2 \leq c\,\|x\|_1$ ist für alle $x \in V$. Wählen wir daher ein M, so daß $\|x_n - x_m\|_1 < \varepsilon/c$ ist für alle $n, m \geq M$, so ist

$$\|x_n - x_m\|_2 \leq c\,\|x_n - x_m\|_1 < c \cdot \frac{\varepsilon}{c} = \varepsilon$$

für alle $n, m \geq M$; die Folge ist also auch bezüglich $\|\cdot\|_2$ eine CAUCHY-Folge. Damit folgt insbesondere, daß V genau dann ein BANACH-Raum bezüglich der Norm $\|\cdot\|_2$ ist, wenn V ein BANACH-Raum bezüglich der dazu äquivalenten Norm $\|\cdot\|_1$ ist.

Speziell für $V = \mathbb{R}^n$, wo *alle* Normen äquivalent sind, reicht es also, die Vollständigkeit bezüglich irgendeiner beliebigen Norm zu beweisen; sie folgt dann automatisch auch für alle anderen Normen.

Lemma: $\mathbb{R}^n$ ist ein BANACH-Raum bezüglich der Maximumsnorm und damit bezüglich jeder beliebigen Norm.

Beweis: $(x_k)_{k\in\mathbb{N}}$ sei eine CAUCHY-Folge von Elementen aus $\mathbb{R}^n$; wir schreiben das n-Tupel $x_k \in \mathbb{R}^n$ als $x_k = (x_{k1}, \ldots, x_{kn})$. Zu jedem $\varepsilon > 0$ gibt es ein $N \in \mathbb{N}$, so daß

$$\|x_k - x_\ell\|_\infty = \max\bigl\{|x_{kj} - x_{\ell j}| \;\big|\; j = 1, \ldots, n\bigr\} < \varepsilon$$

ist für alle $k, \ell \geq N$. Damit ist insbesondere $\bigl|x_{kj} - x_{\ell j}\bigr| < \varepsilon$ für jeden Index j, d.h. die Folgen $(x_{kj})_{k\in\mathbb{N}}$ sind CAUCHY-Folgen reeller Zahlen.

Nach dem CAUCHYschen Konvergenzkriterium konvergiert daher jede dieser Folgen gegen einen Grenzwert $y_j \in \mathbb{R}$. Damit konvergiert die Folge $(x_k)_{k\in\mathbb{N}}$ in $\mathbb{R}^n$ gegen den Punkt $(y_1, \ldots, y_n)$, denn wie wir bereits in §1*b)* gesehen haben, ist Konvergenz bezüglich der Maximumsnorm einfach Konvergenz in jeder Komponente. Somit konvergiert jede CAUCHY-Folge in $\mathbb{R}^n$, und damit ist $\mathbb{R}^n$ vollständig, d.h. ein BANACH-Raum. ∎

Da jeder endlichdimensionale $\mathbb{R}$-Vektorraum V isomorph ist zu einem $\mathbb{R}^n$, sind somit auch alle diese Räume vollständig. Um Vektorräume zu finden, die keine BANACH-Räume sind, müssen wir entweder $\mathbb{Q}$ oder allgemeiner $\mathbb{Q}^n$ betrachten oder aber unendlichdimensionale $\mathbb{R}$-Vektorräume. Im letzten Abschnitt dieses Paragraphen werden wir erste Beispiele von Funktionenräumen betrachten, die teils BANACH-Räume sind, teils auch nicht.

b) Fixpunkte von Abbildungen

Betrachten wir noch einmal das HERON-Verfahren zur näherungsweisen Berechnung von $\sqrt{2}$: Wir starten mit irgendeiner positiven Zahl x_0 und berechnen sukzessive neue Werte

$$x_n = f(x_{n-1}) \quad \text{mit} \quad f(x) = \frac{1}{2}\left(x + \frac{2}{x}\right).$$

Für $x = \sqrt{2}$ ist

$$f\left(\sqrt{2}\right) = \frac{1}{2}\left(\sqrt{2} + \frac{2}{\sqrt{2}}\right) = \frac{1}{2}\left(\sqrt{2} + \sqrt{2}\right) = \sqrt{2}\,;$$

ist umgekehrt x eine positive reelle Zahl mit $f(x) = x$, so ist $f(x) = x$ äquivalent zur Gleichung

$$x = \frac{1}{2}\left(x + \frac{2}{x}\right) \quad \text{oder} \quad \frac{1}{2}\left(\frac{2}{x} - x\right) = 0\,,$$

also ist $x = 2/x$ und somit $x^2 = 2$, was im Positiven nur die Lösung $x = \sqrt{2}$ hat. HERON hat also die Gleichung $x^2 = 2$ umgeschrieben in eine Gleichung $f(x) = x$, und löst sie näherungsweise, indem er auf einen beliebigen Startwert immer wieder die Funktion f anwendet. Lösungen von Gleichungen der Form $f(x) = x$ beschreiben *Fixpunkte* im Sinne der folgenden

Definition: M sei eine Menge und $f\colon M \to M$ eine Abbildung. Ein *Fixpunkt* von f ist ein Element $x \in M$ mit $f(x) = x$.

HERONs iterativer Ansatz funktioniert nicht für jede Funktion: Die Gleichung $x^2 = 2$ ist beispielsweise auch äquivalent zur Gleichung $x = g(x) = 2/x$; wenn wir aber ausgehend von $x_0 = 1$ immer wieder die Funktion g anwenden, pendeln wir nur zwischen den beiden Werten 1 und 2 hin und her, ohne der Wurzel je näher zu kommen.

Ein wichtiges Thema dieses Paragraphen ist die Frage, unter welchen Bedingungen ein iteratives Verfahren wie das von HERON zum Erfolg führt. Wir wollen dieses Problem nicht unter den schwächstmöglichen Voraussetzungen lösen, sondern suchen stattdessen nach einfach zu überprüfenden *hinreichenden* Kriterien. So ist auch die folgende Definition für den eindimensionalen Fall nicht die allgemeinstmögliche:

Definition: $f\colon D \to \mathbb{R}$ sei eine differenzierbare Abbildung auf der offenen Teilmenge $D \subseteq \mathbb{R}$. Ein Punkt $x \in D$ mit $f(x) = x$ heißt *stabiler* oder *anziehender* Fixpunkt von f, wenn $|f'(x)| < 1$ ist; er heißt *instabiler* oder *abstoßender* Fixpunkt, wenn $|f'(x)| > 1$ ist.

(Den Fall $|f'(x)| = 1$ betrachten wir nicht, da er im allgemeinen erheblich schwieriger zu behandeln ist.)

Lemma: Ist x ein anziehender Fixpunkt von f, so gibt es ein $\varepsilon > 0$, so daß für alle $x_0 \in D$ mit $|x - x_0| < \varepsilon$ die durch $x_k = f(x_{k-1})$ definierte Folge gegen x konvergiert. Für einen abstoßenden Fixpunkt dagegen konvergiert diese Folge nur dann gegen x, wenn sie bereits nach endlich vielen Iterationen den Wert$x$ erreicht.

Beweis: Sei zunächst x ein anziehender Fixpunkt. Für alle $h \in \mathbb{R}$ mit $y = x + h \in D$ ist dann

$$f(x + h) = f(x) + f'(x)h + o(h) = x + f'(x)h + o(h)\,.$$

Wir schreiben $|f'(x)| = 1 - 2c$; da $|f'(x)|$ kleiner ist als eins, ist die so definierte Zahl c positiv. Der Fehlerterm $o(h)$ geht schneller gegen Null geht als h; daher gibt es ein $\varepsilon_1 > 0$, so daß $|o(h)| < ch$ für alle h mit $|h| < \varepsilon_1$. Für solche h ist daher

$$\begin{aligned} |f(y) - f(x)| = |f(x + h) - x| = |f'(x)h + o(h)| &\le |f'(x)h| + |o(h)| \\ &< (1 - 2c)\,|h| + c\,|h| = (1 - c)\,|h| \le |h| = |y - x|\ . \end{aligned}$$

Das allein reicht allerdings noch nicht, denn wir wissen nicht, ob $f(y)$ im Definitionsbereich D liegt, so daß wir f auch iterieren können.

Da D offen ist, gibt es aber ein $\varepsilon_2 > 0$, so daß alle $z \in \mathbb{R}^n$ mit $|z - x| < \varepsilon_2$ in D liegen. Nehmen wir nun als ε das Minimum von ε_1 und ε_2, so ist wieder für jedes $y \in D$ mit $|y - x| < \varepsilon$

$$|f(y) - x| \le (1 - c)\,|y - x| < \varepsilon\,,$$

und da $\varepsilon \le \varepsilon_2$ ist, liegt $f(y)$ in$D$.

Starten wir also mit einem x_0, für das $|x - x_0| < \varepsilon$ ist, so folgt induktiv, daß auch alle x_k mit $k \ge 1$ in D liegen, so daß wir die Folge $(x_k)_{k\in\mathbb{N}}$ definieren können; außerdem ist

$$|x - x_k| \le (1 - c)\,\big|x - x_{k-1}\big| \le \cdots \le (1 - c)^k\,|x - x_0| < (1 - c)^k\varepsilon\,.$$

Dies zeigt, daß die Folge der x_k gegen x konvergiert.

Gehen wir allerdings aus von einem abstoßenden Fixpunkt x, so ist $|f'(x)| > 1$; wir schreiben dies als $1 + 2c$ mit einer reellen Zahl $c > 0$. Wieder gibt es ein $\varepsilon > 0$, so daß in der Formel

$$f(x + h) = f(x) + f'(x)h + o(h) = x + f'(x) + o(h)$$

der Betrag von $o(h)$ kleiner ist als c für alle h mit $|h| < \varepsilon$. Für ein $y = x + h$ mit $|h| < \varepsilon$ ist daher

$$\begin{aligned}|f(y) - f(x)| = |f(x + h) - x| = |f'(x)h + o(h)| &\ge |f'(x)h| - |o(h)| \\ &> (1 - 2c)\,|h| + c\,|h| = (1 - c)\,|h| \ge |h| = |y - x|\ .\end{aligned}$$

Nehmen wir nun an, für irgendein x_0 aus D lasse sich die Folge $(x_k)_{k\in\mathbb{N}}$ definieren, und sie konvergiere gegen x. Dann gäbe es ein $N \in \mathbb{N}$, so daß $|x_n - x| < \varepsilon$ wäre für alle $n \ge N$. Für jedes $n \ge N$ wäre daher $|x_n - x| \ge |x_{n+1} - x|$, was für eine gegen x konvergierende Folge nur dann möglich ist, wenn alle $x_n = x$ sind.

∎

Im Mehrdimensionalen ist die Situation *im Prinzip* genauso; der Beweis erfordert allerdings Sätze aus der Linearen Algebra, die nicht allen Hörern bekannt sind. Daher sei das Ergebnis nur kurz skizziert:

Für eine differenzierbare Abbildung $f\colon D \to \mathbb{R}^n$ auf einer offenen Teilmenge $D \subseteq \mathbb{R}^n$ mit Fixpunkt x ist, solange $x + h$ in D liegt,

$$f(x + h) = f(x) + J_f(h) \cdot h + o\big(\|h\|\big) = x + J_f(h) \cdot h + o(\|h\|)$$

und damit $\|f(x+h)-x\| \leq \left\|J_f(h)\cdot h\right\| + \left\|o\big(\|h\|\big)\right\|$.

Den Fehlerterm können wir wieder für hinreichend kleine Norm von h unter jede gewünschte positive Schranke bringen; wir brauchen also in erster Linie eine Schranke für die Norm von $J_f(h)\cdot h$.

Falls $J_f(h)$ eine Diagonalmatrix ist, können wir vorgehen wie im eindimensionalen Fall: Die i-te Komponente des Vektors h wird mit dem i-ten Diagonaleintrag der Matrix multipliziert; falls dieser einen Betrag kleiner eins hat, konvergiert die Folge der iterierten Produkte zumindest in der i-ten Komponente gegen Null. Die Folge der x_k mit $x_k = f(x_{k-1})$ konvergiert also für einen Anfangswert x_0 hinreichend nahe bei x genau dann gegen x, wenn alle Diagonaleinträge Beträge kleiner als eins haben.

Leider ist $J_f(h)$ nur in den seltensten Fällen eine Diagonalmatrix. Für die Frage, ob eine Folge von Vektoren x_k gegen einen Vektor x konvergieren, ist es aber gleichgültig, in welcher Basis wir die Vektoren ausdrücken; falls es also eine Basis gibt, bezüglich derer $J_f(h)$ eine Diagonalmatrix ist, können wir wie oben argumentieren. Dabei spielt es keine Rolle, ob wir eine solche Basis für $\mathbb{R}^n$ oder nur für $\mathbb{C}^n$ finden können.

Die Einträge der Diagonalmatrix sind bekanntlich gerade die Eigenwerte von $J_f(h)$; die Folge der Iterierten konvergiert also, falls die Matrix $J_f(h)$ diagonalisierbar ist und alle ihre Eigenwerte Beträge kleiner eins haben.

Falls $J_f(h)$ nicht diagonalisierbar ist, läßt sich die Matrix als eine Summe $J_f(h) = D + N$ schreiben mit einer diagonalisierbaren Matrix D und einer Matrix N, deren Potenzen ab einem gewissen Exponenten $r \leq n$ gleich der Nullmatrix sind. Außerdem kommutieren D und N, d.h. $DN = ND$. Deshalb können wir auf diese speziellen Matrizen den binomischen Lehrsatz anwenden und erhalten

$$J_f(h)^m = (D+N)^m = \sum_{k=0}^{m} \binom{m}{k} D^{m-k} N^k \,.$$

Für $k \geq r$ verschwindet N^k; für $m \geq r$ ist also

$$J_f(h)^m = (D+N)^m = \sum_{k=0}^{r-1} \binom{m}{k} D^{m-k} N^k = D^{m-r} \sum_{k=0}^{r-1} \binom{m}{k} D^{r-k} N^k \,.$$

Da r eine feste, von m unabhängige Zahl ist, wird das Verhalten von $J_f(h)^m \cdot h$ für $m \to \infty$ daher wieder von den Eigenwerten von $J_f(x)$ kontrolliert; auch hier konvergiert also die Folge der Iterierten, falls alle Eigenwerte von $J_f(h)$ einen Betrag kleiner eins haben. Deshalb definieren wir

Definition: $f\colon D \to \mathbb{R}^n$ sei eine differenzierbare Abbildung auf der offenen Teilmenge $D \subseteq \mathbb{R}^n$. Ein Punkt $x \in D$ mit $f(x) = x$ heißt *stabiler* oder *anziehender* Fixpunkt von f, wenn alle Eigenwerte der JACOBI-Matrix $J_f(x)$ einen Betrag kleiner eins haben; er heißt *instabiler* oder *abstoßender* Fixpunkt, wenn mindestens ein Eigenwert einen größeren Betrag als eins hat.

Damit läßt sich, mit praktisch demselben Beweis, das obige Lemma auch für höhere Dimensionen zeigen.

c) Die Lorenz-Gleichungen

Fixpunkte sind zwar nur einzelne isolierte Punkte; sie können aber doch oft erstaunlich viel über eine Funktion aussagen. Als Beispiel dazu betrachten wir die sogenannten LORENZ-Gleichungen

$$\dot{x}(t) = p\big(y(t) - x(t)\big)$$
$$\dot{y}(t) = rx(t) - y(t) - x(t)z(t)$$
$$\dot{z}(t) = -bz(t) + x(t)y(t)$$

Der amerikanische Mathematiker und Meteorologe EDWARD LORENZ stellte sie auf als eine extreme Vereinfachung der NAVIER-STOKES-Gleichungen für die Dynamik der Atmosphäre. Die drei Funktionen $x(t), y(t)$ und $z(t)$ haben keine direkte physikalische Interpretation, sondern hängen mit niederfrequenten FOURIER-Moden von Lösungen der NAVIER-STOKES-Gleichungen zusammen; p, q und r sind Parameter, die laut LORENZ für die atmosphärische Konvektion ungefähr bei $p = 10$, $r = 28$ und $b = \frac{8}{3}$ liegen sollten.

Gesucht sind also Funktionen $x(t), y(t), z(t)$, für deren Ableitungen $\dot{x}(t), \dot{y}(t)$ und $\dot{z}(t)$ die obigen Gleichungen gelten. (Ableitungen nach der Zeit werden oft durch einen Punkt statt einen Strich bezeichnet.) Die Lösungen dieses sogenannten Differentialgleichungssystems können

nicht in geschlossener Form angegeben werden, man kann sie aber mit numerischen Methoden näherungsweise bestimmen, was LORENZ auch tat.

EDWARD NORTON LORENZ (1917–2008) stammt aus dem US-Bundesstaat Connecticut; er studierte Mathematik am Dartmouth College (A.B. 1938) und in Harvard (M.A. 1940). Nach seinem Kriegsdienst ging er ans MIT, wo er 1948 über Meteorologie promovierte. Sowohl dem MIT, wo er 1987 als Professor emeritiert wurde, als auch der Meteorologie blieb er fortan treu. Zu seinen vielen Auszeichnungen gehört unter anderem der Kyoto-Preis von 1991, der wohl höchstdotierte Wissenschaftspreis.

Die einfachste Art, eine Lösungskurve näherungsweise zu bestimmen, geht auf EULER zurück: Man wählt einen Startzeitpunkt t_0 mit zugehörigen Startwerten $x(t_0) = x_0$, $y(t_0) = y_0$ und $z(t_0) = z_0$; für eine geeignete Schrittweite h bestimmt man daraus näherungsweise nacheinander die Funktionswerte an den Stellen $t_0 + nh$ für $n \in \mathbb{N}$ durch die Formeln

$$x(t+h) \approx x(t) + h\dot{x}(t) = x(t) + hp\big(y(t) - x(t)\big)$$
$$y(t+h) \approx y(t) + h\dot{y}(t) = y(t) + h\big(rx(t) - y(t) - x(t)z(t)\big)$$
$$z(t+h) \approx z(t) + h\dot{z}(t) = (1 - hb)z(t) + hx(t)y(t)$$

Der programmierbare elektromechanische Rechner, mit dem LORENZ arbeitete, stürzte im Laufe der Rechnungen immer wieder ab; um dann nicht wieder ganz von vorne anfangen zu müssen, notierte LORENZ von Zeit zu Zeit Zwischenwerte, so daß er gegebenenfalls die Iteration dort neu beginnen lassen konnte. Für diese Notitzen begnügte er sich mit dreistelliger Genauigkeit. Zu seinem Erstaunen stellte er fest, daß eine an einem solchen späteren Zeitpunkt wiederaufgenommene Rechnung schon nach wenigen Iterationen zu ganz anderen Ergebnissen führte als eine in einem Stück durchgeführte; durch genauere Untersuchungen fand er heraus, daß selbst kleinste Änderungen bei den Startwerten zu dramatischen Änderungen im weiteren Verlauf der Rechnung führen.

Falls das gleiche Phänomen auch in der wirklichen Atmosphäre auftritt, können also minimale Veränderungen etwa des Luftdrucks oder der Temperatur auf längere Sicht zu einer dramatisch anderen Entwicklung

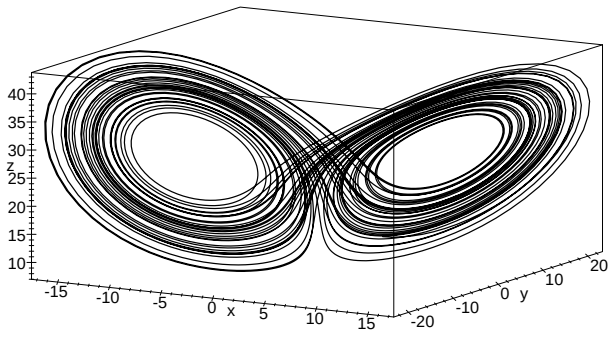
des Wetters führen – eine Idee, die vielen Meteorologen damals als zu phantastisch erschien um ernstgenommen zu werden: Am 22. Januar 1963 berichtete LORENZ vor der New York Academy of Sciences über seine Ergebnisse *(Trans. N.Y. Acad. Sci.* **25** *(1963), 409–432)* und schloß seinen Vortrag mit den Worten:

Als die Instabilität eines gleichförmigen Flusses gegenüber infinitesimalen Störungen erstmals als Erklärung für das Auftreten von Zyklonen und Antizyklonen in der Atmosphäre vorgeschlagen wurde, war diese Idee nicht allgemein akzeptiert. Ein Meteorologe bemerkte, daß, falls die Theorie korrekt wäre, ein Flügelschlag einer Möwe ausreichen würde, um die Entwicklung des Wetters für immer zu verändern. Die Kontroverse ist noch nicht entschieden, aber die neueste Evidenz scheint für die Möwen zu sprechen.

Inzwischen ist der Sieg der Möwen bekanntlich allgemein anerkannt; man fordert sogar nicht einmal mehr den relativ kräftigen Flügelschlag einer Möwe, um das Wetter permanent zu verändern: Im Dezember 1972 hielt LORENZ vor der American Association for the Advancement of Sciences in Washington, DC, einen Vortrag mit dem Titel *Predictability: Does the Flap of a Butterfly's Wings in Brazil set off a Tornado in Texas*, und seitdem geht das Wort vom *Schmetterlingseffekt* um die Welt.

Auch das Wort *Chaos* wird heute meist auf diese Weise definiert: Kleinste Änderungen bei den Anfangsbedingungen führen zu dramatischen Veränderungen des Langzeitverhaltens.

Chaos heißt nun allerdings nicht, daß wir dann überhaupt nichts über das Verhalten der Lösungen aussagen können. Wenn wir numerisch rechnen, erhalten wir erstaunlicherweise unabhängig von den Startwerten immer ein Bild, das ungefähr so aussieht, wie das unten abgedruckte.



Da wir mit endlicher Genauigkeit rechnen, können wir sicher sein, daß unsere näherungsweise berechneten Lösungskurven wegen der unvermeidbaren Rundungsfehler quantitativ schon kurz nach der Startzeit nichts mehr mit den exakten Lösungskurven zu tun haben; sie geben aber offensichtlich das qualitative Verhalten recht gut wieder.

Um das zu verstehen, betrachten wir die Fixpunkte, d.h. wir suchen nach konstanten Lösungen des Systems. Da die Ableitung einer konstanten Funktion verschwindet, sind die Fixpunkte Lösungen des Gleichungssystems

$$\begin{aligned} 0 &= p(y-x) \\ 0 &= rx - y - xz \\ 0 &= -bz + xy \end{aligned}$$

Wenn wir den uninteressanten Fall $p = 0$ ausschließen, folgt aus der ersten Gleichung, daß für jeden Fixpunkt $x = y$ sein muß. Falls beide verschwinden, ist nach der dritten Gleichung auch $z = 0$; als ersten Fixpunkt erhalten wir also den Nullpunkt.

Im Fall $x \neq 0$ können wir y in der zweiten Gleichung durch x ersetzen und dann durch x dividieren; dies ergibt die z-Koordinate

$$z = r - 1\,.$$

Damit zeigt die dritte Gleichung, daß es für $r \neq 1$ noch zwei weitere Fixpunkte gibt mit

$$x = y = \pm\sqrt{b(r-1)} \quad \text{und} \quad z = r - 1\,.$$

In unserer Terminologie sind die drei gefundenen Punkte Fixpunkte der Abbildung $F\colon \mathbb{R}^3 \to \mathbb{R}^3$ mit

$$F(x,y,z) = \begin{pmatrix} x + hp(y-x) \\ y + h(rx - y - xz) \\ (1-hb)z + hxy \end{pmatrix};$$

ihre JACOBI-Matrix ist

$$J_F(x,y,z) = \begin{pmatrix} 1-hp & hp & 0 \\ h(r-z) & 1-h & -hx \\ hy & hx & 1-hb \end{pmatrix}.$$

Um Aussagen über die Stabilität zu machen, müssen wir die Eigenwerte dieser Matrix bestimmen. Im Nullpunkt haben wir

$$J_F(0,0,0) = \begin{pmatrix} 1-10h & 10h & 0 \\ 28h & 1-h & 0 \\ 0 & 0 & 1-\frac{8}{3}h \end{pmatrix};$$

die Eigenwerte dieser Matrix sind

$$1-\frac{8}{3}h, \qquad 1+\frac{1}{2}(\sqrt{1201}-11)h \quad \text{und} \quad 1-\frac{1}{2}(\sqrt{1201}+11)h\,.$$

Für eine kleine positive Schrittweite h ist somit der mittlere der drei Eigenwerte größer als eins, die beiden anderen sind kleiner. Der Nullpunkt ist daher kein stabiler Fixpunkt; er stößt ab in Richtung des Eigenvektors zum zweiten Eigenwert.

Für die beiden anderen Fixpunkte erhalten wir die JACOBI-Matrizen

$$J_F(\pm\sqrt{b(r-1)}, \pm\sqrt{b(r-1)}, r-1) = \begin{pmatrix} 1-10h & 10h & 0 \\ h & 1-h & \pm 6h\sqrt{2} \\ \pm 6h\sqrt{2} & \pm 6h\sqrt{2} & 1-\frac{8}{3}h \end{pmatrix},$$

deren Eigenwerte, wenn wir sie allgemein ausrechnen, recht grausame Ausdrücke sind. Wir setzen daher nicht nur p, b und r auf die von LORENZ angegebenen Werte, sondern legen auch noch die Schrittweite h fest. Für $h = 0{,}01$ bekommen wir die Eigenwerte

$$\lambda_1 \approx 0{,}8614542208 \quad \text{und} \quad \lambda_{2/3} \approx 1{,}000939556 \pm 0{,}1019450522i\,.$$

Wir haben somit einen reellen Eigenwert vom Betrag kleiner eins sowie zwei konjugiert komplexe vom Betrag größer eins. Keiner der drei Fixpunkte ist also stabil.

Trotzdem sagen uns diese Eigenwerte einiges über das Verhalten der Lösungskurven: Der Eigenwert λ_1 sorgt dafür, daß eine Lösungskurve, wenn sie erst einmal in der Nähe des jeweiligen Fixpunkts ist, in Richtung derjenigen Ebene durch den Fixpunkt gedrückt wird, die von den Eigenvektoren zu λ_2 und λ_3 aufgespannt wird, während λ_2 und λ_3 dafür sorgen, daß sie innerhalb dieser Ebene spiralförmig vom Fixpunkt weggetrieben werden. Sind sie erst einmal weit genug weg vom Fixpunkt, verliert dieser seinen Einfluß, sie können sich also wieder von der Ebene entfernen. Dadurch können sie in den Einflußbereich des anderen Fixpunkts kommen, wo im wesentlichen das Gleiche passiert. Wir

können uns also die Lösungskurven des LORENZ-Systems so vorstellen, daß hier zwei abstoßende Fixpunkte miteinander Schleuderball spielen, was auch gut zur oben abgebildeten Kurve paßt.

d) Das Newton-Verfahren

Nicht nur bei Extremwertproblemen, egal ob mit oder ohne Nebenbedingungen, ist es oft notwendig, die Nullstellen einer nichtlinearen Gleichung oder eines nichtlinearen Gleichungssystems zu bestimmen. Nur in sehr speziellen Fällen können diese Nullstellen exakt berechnet werden; meist muß man sich mit Näherungslösungen zufrieden geben.

Die numerische Mathematik kennt daher zahlreiche Methoden zur näherungsweisen Berechnung von Nullstellen; alle haben sowohl Stärken als auch Schwächen.

Das hier betrachtete NEWTON-Verfahren wird gerne verwendet bei differenzierbaren Funktionen, deren Ableitung sich einfach berechnen läßt, also beispielsweise bei Polynomen. Es wurde 1669 von ISAAC NEWTON vorgeschlagen und unabhängig davon 1690 von JOSEPH RAPHSON neu entdeckt und in der heute gebräuchlichen Form publiziert. Man bezeichnet es daher oft auch als Verfahren von NEWTON-RAPHSON. Wie viele numerische Verfahren ist es ein Iterationsverfahren; solche Verfahren haben den Vorteil, daß sie Rundungsfehler, die in einem Iterationsschritt entstehen, in den Folgeschritten im allgemeinen nicht vergrößern, sondern verkleinern.

SIR ISAAC NEWTON wurde gemäß dem damals noch in England geltenden Julianischen Kalender am 25. Dezember 1642 geboren. Nach dem in den meisten katholischen Staaten bereits eingeführten Gregorianischen Kalender war dies der 4. Januar 1643. Er studierte ab 1661 an der Universität Cambridge, wo er 1669 Professor wurde. Dort entwickelte er die Infinitesimalrechnung, die er 1671 in seinem Buch *De Methodis Serierum et Fluxionum* beschrieb, arbeitete über Optik, wo er unter anderem dünne Schichten und Beugungsphänomene untersuchte (NEWTONsche Ringe), entdeckte seine Bewegungsgesetze und das Gravitationsgesetz, veröffentlicht 1687 in seinem Buch *Philosophiae naturalis principia mathematica,* das von vielen als bedeutendstes wis-

senschaftliches Buch aller Zeiten angesehen wird. Nach zwei Nervenzusammenbrüchen ging er 1693 nach London, wo er die königliche Münze leitete. Er starb am 31. März 1727.

Über die Biographie von JOSEPH RAPHSON (1648–1715) ist sehr viel weniger bekannt. Auch er studierte in Cambridge, wo er 1692 seinen M.A. erhielt; bereits 1690 veröffentlichte er sein Buch *Analysis Aequationum universalis,* das seine Version des NEWTON-Verfahrens enthält, und wurde 1691 Mitglied der Royal Society. Spätere Bücher beschäftigen sich außer mit Mathematik auch mit theologischen und naturphilosophischen Fragen.

Gegeben sei eine differenzierbare Funktion $f\colon \mathbb{R} \to \mathbb{R}$; wir suchen eine Nullstelle x von f. Dem Thema dieses Paragraphen entsprechend wollen wir x als Fixpunkt einer Abbildung interpretieren und iterativ berechnen.

Der einfachste denkbare Ansatz besteht darin, daß wir die Gleichung $f(x) = 0$ umschreiben als $x = x + f(x)$. Testet man diesen Ansatz mit einfachen Funktionen $f(x)$, so merkt man schnell, daß er nur selten zu einem brauchbaren Ergebnis führt.

Das NEWTON-Verfahren geht aus von folgender Beobachtung: Ist x_0 eine *einfache* Nullstelle von f, d.h. $f'(x) \neq 0$, so schneidet die Tangente an die Kurve $y = f(x)$ im Punkt mit x-Koordinate x_0 die x-Achse an der Stelle $x = x_0$.

Für ein beliebiges x_0 aus dem Definitionsbereich von f, in dem $f'(x_0)$ nicht verschwindet, hat die Tangente im Punkt $\big(x_0, f(x_0)\big)$ die Gleichung $y = f(x_0) + f'(x_0)(x - x_0)$ und schneidet daher die x-Achse im Punkt

$$x = x_0 - \frac{f(x_0}{f'(x_0)}\,,$$

der in der Tat genau dann mit x_0 übereinstimmt, wenn $f(x_0)$ verschwindet. Zur iterativen Bestimmung einer Nullstelle können wir also versuchen, mit irgendeinem Startwert $x_0 \in D$ anzufangen und weitere Näherungswerte zu bestimmen durch die Vorschrift

$$x_n = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})} \quad \text{für alle } n \in \mathbb{N}\,.$$

Als erstes Beispiel betrachten wir die Funktion $f(x) = x^2 - a$, deren Nullstellen die Quadratwurzeln von a sind. Hier ist $f'(x) = 2x$, für

$x_i \neq 0$ haben wir also die Iterationsvorschrift

$$x_n = x_{n-1} - \frac{x_{n-1}^2 - a}{2x_{n-1}} = x_{n-1} - \frac{x_{n-1}}{2} + \frac{a}{2x_{n-1}} = \frac{1}{2}\left(x_{n-1} + \frac{a}{x_{n-1}}\right),$$

die HERON bereits rund sechzehn Jahrhunderte vor NEWTON benutzte, und von der wir gesehen haben, daß sie schnell gute Ergebnisse liefert.

Auch bei komplizierteren Polynomen hat sich das NEWTON-Verfahren in der Praxis sehr bewährt. Um zu verstehen, warum das so ist, betrachten wir die Funktion

$$\varphi(x) = x - \frac{f(x)}{f'(x)},$$

die für alle x mit $f'(x) \neq 0$ definiert ist und die uns zu einem Iterationswert x_n den Folgewert x_{n+1} liefert. Offensichtlich ist z genau dann eine einfache Nullstelle von f, wenn $\varphi(z) = z$ ist; die einfachen Nullstellen von f sind also genau die Fixpunkte von φ.

Angenommen, wir haben uns einem solchen Fixpunkt bis auf die Distanz h genähert, d.h. wir haben ein $x_n = z + h$. Wir wollen abschätzen, wie weit dann $x_{n+1} = \varphi(x_i)$ von z entfernt ist.

Falls h klein ist, können wir auch φ ohne großen Fehler durch seine Linearisierung ersetzen:

$$x_{n+1} = \varphi(z + h) \approx \varphi(z) + h\varphi'(z) = z + h\varphi'(z).$$

Die Ableitung von φ können wir leicht nach der Quotientenregel berechnen:

$$\varphi(x) = x - \frac{f(x)}{f'(x)} \Longrightarrow \varphi'(x) = 1 - \frac{f'(x)^2 - f(x)f''(x)}{f'(x)^2}$$
$$= 1 - 1 + \frac{f(x)f''(x)}{f'(x)^2} = \frac{f(x)f''(x)}{f'(x)^2}.$$

Somit ist

$$x_{n+1} \approx z + h\frac{f(z)f''(z)}{f'(z)^2} = z,$$

da $f(z)$ verschwindet.

Dieses Ergebnis war, wenn man ein bißchen nachdenkt, natürlich zu erwarten; es ist aber völlig nutzlos, um den Abstand zwischen x_{i+1} und z

zu bestimmen. Wenn wir ein nützliches Ergebnis erhalten wollen, dürfen wir uns also nicht auf eine lineare Approximation beschränken, sondern müssen zumindest auch noch den quadratischen Term berücksichtigen.

Wir gehen daher aus von der Approximation

$$\varphi(x+h) \approx \varphi(x) + h\varphi'(x) + \tfrac{1}{2}h^2\varphi''(x)\,,$$

$\varphi''(x)$ ist die Ableitung von $\varphi'(x) = \frac{f(x)f''(x)}{f'(x)^2}$, also ist nach der Quotientenregel

$$\varphi''(x) = \frac{f'(x)^2\big(f'(x)f''(x) + f(x)f'''(x)\big)}{f'(x)^4}\,.$$

Speziell für $x = z$, wo $f(z)$ verschwindet und $\varphi(z) = z$ ist, erhalten wir die Abschätzung

$$\varphi''(z) = \frac{f''(z)}{f'(z)} \quad \text{und} \quad \varphi(z+h) \approx z + \frac{h^2}{2}\frac{f''(z)}{f'(z)}\,.$$

Der Abstand des neuen Iterationswert zur Nullstelle z ist also für kleine Werte von h bis auf einen nur von z abhängigen Vorfaktor gleich dem Quadrat des alten und verkleinert sich somit bei kleinen Werten von h sehr schnell.

Leider ist diese Aussage nicht so konkret, daß wir für irgendeinen vorgegebenen Startwert entscheiden können, ob und gegebenenfalls wohin das NEWTON-Verfahren konvergiert, denn wir wissen nicht, wann der Abstand h „klein“ ist oder wird. Betrachten wir dazu als Beispiel das Polynom $f(x) = x^3 - 5x = x(x^2 - 5)$; seine Nullstellen sind offensichtlich $x = 0$ und $x = \pm\sqrt{5}$. Die zu iterierende Funktion ist hier

$$\varphi(x) = x - \frac{f(x)}{f'(x)} = x - \frac{x^3 - 5x}{3x^2 - 5}\,.$$

Für $x_0 = 1$ ist daher

$$x_1 = \varphi(1) = 1 - \frac{-4}{-2} = -1 \quad \text{und} \quad x_2 = \varphi(-1) = -1 - \frac{4}{-2} = 1\,.$$

Damit ist klar, daß x_n für alle geraden n gleich eins ist und für die ungeraden -1; mit Startwert $x_0 = 1$ (oder -1) bekommen wir also nie ein nützliches Ergebnis.

Die folgende Tabelle zeigt, was passiert, wenn wir x_0 leicht vergrößern. (Die Werte der x_i sind zwar nur mit fünf geltenden Ziffern angegeben,

die Iterationen wurden aber mit hundertstelliger Genauigkeit gerechnet.)

	x_0	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
$1+10^{-1}$	-1,9431	-2,3191	-2,2403	-2,2361	-2,2361	-2,2361	-2,2361	-2,2361	-2,2361	-2,2361	-2,2361	-2,2361	-2,2361
$1+10^{-2}$	-1,0623	1,4853	4,0498	3,0053	2,4569	2,2627	2,2365	2,2361	2,2361	2,2361	2,2361	2,2361	2,2361
$1+10^{-3}$	-1,0060	1,0370	-1,2570	15,310	10,280	6,9631	4,8073	3,4540	2,6766	2,3254	2,2410	2,2361	
$1+10^{-4}$	-1,0006	1,0036	-1,0220	1,1435	-2,7749	-2,3610	-2,2453	-2,2361	-2,2361	-2,2361	-2,2361	-2,2361	
$1+10^{-5}$	-1,0001	1,0004	-1,0022	1,0131	-1,0826	1,7096	2,6521	2,3171	2,2401	2,2361	2,2361	2,2361	
$1+10^{-6}$	-1,0000	1,0000	-1,0002	1,0013	-1,0078	1,0483	-1,3532	-10,050	-6,8125	-4,7108	-3,3956	-2,6462	
$1+10^{-7}$	-1,0000	1,0000	-1,0000	1,0001	-1,0008	1,0047	-1,0286	1,1920	-4,5915	-3,3238	-2,6095	-2,3035	
$1+10^{-8}$	-1,0000	1,0000	-1,0000	1,0000	-1,0001	1,0005	-1,0028	1,0170	-1,1090	2,0814	2,2552	2,2363	

Das Verfahren konvergiert offensichtlich gegen eine der beiden Nullstellen $\sqrt{5}$ oder $-\sqrt{5}$, wobei keine Regel erkennbar ist, wann es gegen welche der beiden konvergiert. Wenn wir dieselbe Rechnung ausführen für die Startwerte $1 - 10^{-i}$, erhalten wir ein langweiligeres Ergebnis: Nun konvergiert das Verfahren stets gegen die dritte Nullstelle Null.

Der Startwert $x_0 = 1$ liegt also in der Nähe der Einzugsbereiche aller drei Nullstellen, was verständlich macht, daß dort Probleme auftreten.

Ein ähnliches Problem bekommen wir, wenn wir das NEWTON-Verfahren anwenden zur Nullstellenbestimmung des Polynoms $f(x) = x^2 + 2$: Hier ist

$$x_{n+1} = x_n - \frac{x_n^2 + 2}{2x_n} = \frac{1}{2}\left(x_n - \frac{2}{x_n}\right)$$

für reelles x_n natürlich auch wieder reell, die Folge der x_n kann also für keinen reellen Startwert x_0 gegen eine der Nullstellen $\pm\sqrt{-2}$ konvergieren. Für $x_0 = 1$ beispielsweise erhalten wir die Folge

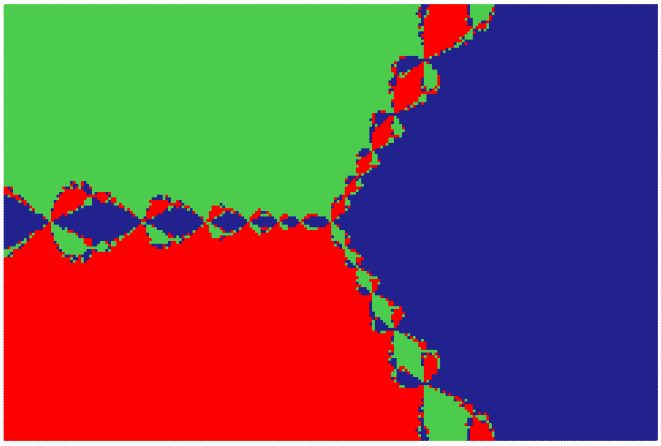
$$-\tfrac{1}{2},\ 1\tfrac{3}{4},\ 0{,}30357,\ -3{,}14233,\ -1{,}25293,\ 0{,}17166,\ -5{,}73953,\ \ldots\ .$$

Wenn wir allerdings mit $x_0 = i$ oder z.B. mit $x_0 = 1+2i$ beginnen, haben wir bereits nach wenigen Iterationen ein Ergebnis mit zehn korrekten Nachkommastellen sowohl im Realteil als auch in Imaginärteil.

Das folgende Bild zeigt das Verhalten des NEWTON-Verfahrens im Komplexen anhand des Polynoms $f(x) = x^3 - 1$, von dem wir aus der Analysis I wissen, daß es die drei komplexen Nullstellen

$$1, \quad \rho = -\frac{1}{2} + \frac{\sqrt{3}}{2}i \quad \text{und} \quad \bar{\rho} = -\frac{1}{2} - \frac{\sqrt{3}}{2}i$$

hat. Bei allen blau gezeichneten Startwerten konvergiert das Verfahren gegen eins, für die grünen gegen ρ und für die roten gegen $\overline{\rho}$. In der Umgebung jeder Nullstelle haben alle Punkte deren Farbe, an den Grenzen gibt es eine komplizierte (fraktale) Struktur. Treffpunkt der drei Grenzgebiete ist der Nullpunkt, der nicht als Startpunkt genommen werden kann, da $f'(x) = 3x^2$ dort verschwindet.



Hauptanwendung des NEWTON-Verfahrens ist allerdings die Bestimmung reeller Nullstellen. Wie wir gesehen haben, gibt es auch hier keine Garantie, daß es mit einem vorgegebenen Startwert wirklich gegen eine Nullstelle konvergiert; es gibt allerdings eine ganze Reihe von Untersuchungen, die auf hinreichende Kriterien führen. Ein einfaches Beispiel ist das folgende:

f sei ein Polynom und $a \leq b$ seien zwei reelle Zahlen, für die $f(a) \cdot f(b) < 0$ ist. Dann haben $f(a)$ und $f(b)$ verschiedene Vorzeichen, also muß es zwischen a und b mindestens eine Nullstelle von f geben. Weiterhin sei $f'(x)$ entweder positiv für alle x mit $a \leq x \leq b$ oder aber negativ für alle solche x. Damit steigt oder fällt der Graph von $f(x)$ zwischen a und b monoton, so daß es dort nur *eine* Nullstelle geben kann. Außerdem folgt, daß das Maximum von $|f(x)|$ im Intervall $[a, b]$ in einem der beiden Endpunkte angenommen wird; diesen Endpunkt bezeichnen wir mit c. Um zu verhindern, daß die Tangentensteigungen zu sehr schwanken und uns die Iterationen aus dem Intervall $[a, b]$ hinausführen, verlangen wir außerdem noch, daß $f''(x)$ dort entweder überall nichtnegativ ist (d.h. der Graph ist konvex) oder überall

nichtpositiv (Graph konkav), und wir fordern, daß

$$\left| \frac{f(c)}{f'(c)} \right| \leq b - a$$

ist; daraus folgt, daß die Tangente im Punkt $\big(c, f(x)\big)$ die x-Achse im Intervall $[a, b]$ schneidet. Dann zeigt eine nicht sehr aufwendige Rechnung, daß das NEWTON-Verfahren für jeden Startwert x_0 zwischen a und b gegen die eindeutig bestimmte Nullstelle in $[a, b]$ konvergiert.

Ähnliche und auch sehr viel allgemeinere Aussagen findet man (mit Beweisen) in praktisch jedem Lehrbuch der Numerischen Mathematik; allgemein läßt sich sagen, daß sich das NEWTON-Verfahren in der Praxis fast immer sehr gut verhält, daß sich aber wirklich allgemeine theoretische Aussagen nur schwer beweisen lassen.

Selbstverständlich läßt sich das NEWTON-Verfahren auch auf mehrdimensionale Probleme anwenden: Wenn wir eine differenzierbare Funktion $f\colon D \to \mathbb{R}^n$ haben mit $D \subseteq \mathbb{R}^n$, können wir sie in der Nähe eines Punktes $x_0 \in D$ annähern durch die lineare Funktion

$$\ell(x) = f(x_0) + J_f(x_0) \cdot (x - x_0)\,.$$

Diese Funktion verschwindet genau dann, wenn x eine Lösung des linearen Gleichungssystems

$$J_f(x_0) \cdot x = J_f(x_0) \cdot x_0 - f(x_0)$$

ist. Falls die JACOBI-Matrix $J_f(x_0)$ invertierbar ist, hat dieses Gleichungssystem genau eine Lösung, nämlich $x = x_0 - J_f(x_0)^{-1} f(x_0)$. Daher können wir auch hier eine Iteration definieren durch

$$x_n = x_{n-1} - J_f(x_n)^{-1} f(x_{n-1})$$

– immer vorausgesetzt, die Matrizen $J_f(x_n)$ werden nie singulär.

e) Der Banachsche Fixpunktsatz

In den beiden letzten Abschnitten hatten wir jeweils Aussagen darüber, daß eine Folge $(x_k)_{k\in\mathbb{N}}$ mit $x_k = f(x_{k-1})$ unter gewissen Voraussetzungen gegen einen vorgegebenen Fixpunkt x von f konvergiert, wenn der Startwert x_0 hinreichend nahe bei x liegt. Es sagt uns aber weder, *wie* nahe das in einem konkreten Fall sein muß, noch sagt es uns, ob es

überhaupt einen Fixpunkt gibt. In diesem Abschnitt geht es um einen Satz, der sowohl die Existenz eines Fixpunkts als auch die Konvergenz dorthin unabhängig von Startwert garantiert. Die Voraussetzungen sind natürlich deutlich stärker, was die Anwendbarkeit etwas einschränkt; trotzdem ist der Satz von zentraler Bedeutung sowohl für die reine als auch die angewandte Mathematik. Wir werden ihn daher auch relativ allgemein formulieren und beweisen; auch wenn er uns im Augenblick vor allem für Funktionen auf Teilmengen des $\mathbb{R}^n$ interessiert.

Definition: V sei ein normierter Vektorraum und $X \subseteq V$. Eine Abbildung $f\colon X \to V$ heißt *kontrahierend*, wenn es eine reelle Zahl $q < 1$ gibt, so daß für alle $x, y \in X$ gilt: $\|f(y) - f(x)\| \leq q \|y - x\|$.

Banachscher Fixpunktsatz: V sei ein BANACH-Raum, $X \subseteq V$ eine abgeschlossene Teilmenge, und $f\colon X \to V$ eine kontrahierende Abbildung mit $f(X) \subseteq X$. Dann hat f genau einen Fixpunkt $x^* \in X$, und für jedes $x_0 \in X$ konvergiert die Folge $(x_n)_{n\in\mathbb{N}}$ mit $x_n = f(x_{n-1})$ gegen x^*.

Beweis: $q < 1$ sei die Konstante, für die $\|f(y) - f(x)\| \leq q \|y - x\|$ ist für alle $x, y \in X$. Wir zeigen als erstes, daß es *höchstens* einen Fixpunkt gibt: Ist $x^* = f(x^*)$ und $y^* = f(y^*)$, so ist

$$\|y^* - x^*\| = \|f(x^*) - f(y^*)\| \leq q \|x^* - y^*\| \ .$$

Das ist aber nur möglich, wenn $\|y^* - x^*\| = 0$ ist, also $x^* = y^*$.

Als nächstes wollen wir sehen, daß die Folge der x_n für jeden Startwert eine CAUCHY-Folge ist. Falls $f(x_0) = x_0$ ist, haben wir eine konstante Folge, und alles ist klar. Andernfalls müssen wir Differenzen von Folgengliedern abschätzen; sei also n eine natürliche Zahl und $m = n + k$ für ein $k \in \mathbb{N}$. Dann ist nach der Dreiecksungleichung

$$\begin{aligned} \|x_m - x_n\| = \|x_{n+k} - x_n\| &= \left\| \sum_{j=0}^{k-1} (x_{n+j+1} - x_{n+j}) \right\| \\ &\leq \sum_{j=1}^{k} \left\| x_{n+j+1} - x_{n+j} \right\| \ . \end{aligned}$$

Da f kontrahierend ist, folgt weiter, daß für jedes $\ell \in \mathbb{N}$ gilt

$$\begin{aligned}\|x_{\ell+1} - x_\ell\| &\leq q \left\|x_\ell - x_{\ell-1}\right\| \leq q^2 \left\|x_{\ell-1} - x_{\ell-2}\right\| \leq \cdots \\ &\leq q^\ell \|x_1 - x_0\| \,.\end{aligned}$$

Somit ist nach der Summenformel für die geometrische Reihe

$$\begin{aligned}\|x_m - x_n\| &\leq \sum_{j=0}^{k-1} q^{n+j} \|x_1 - x_0\| = \frac{q^n - q^m}{1-q} \|x_1 - x_0\| \\ &\leq q^n \frac{\|x_1 - x_0\|}{1-q} \,.\end{aligned}$$

Wenn wir uns ein $\varepsilon > 0$ vorgeben, ist somit $\|x_m - x_n\| < \varepsilon$, falls

$$q^n \frac{\|x_1 - x_0\|}{1-q} < \varepsilon \quad \text{oder} \quad q^n < \frac{(1-q)\varepsilon}{\|x_1 - x_0\|} \,.$$

Da rechts eine positive Zahl steht und die Folge der Potenzen von q eine Nullfolge ist, gibt es ein $N \in \mathbb{N}$, so daß diese Ungleichung für alle $n \geq N$ erfüllt ist; $(x_n)_{n\in\mathbb{N}}$ ist also eine CAUCHY-Folge.

Da V als BANACH-Raum vollständig ist, konvergiert diese Folge gegen einen Grenzwert $x^* \in V$. Da alle x_n in der abgeschlossenen Menge X liegen, liegt auch der Grenzwert dort, d.h. $x^* \in V$. Wir wollen uns überlegen, daß x^* ein Fixpunkt von f ist:

Für jedes $\varepsilon > 0$ gibt es ein $M \in \mathbb{N}$, so daß $\|x^* - x_n\| < \frac{1}{3}\varepsilon$ für alle $n \geq M$. Damit ist auch

$$\|f(x^*) - f(x_n)\| \leq q \|x^* - x_n\| \leq \frac{q\varepsilon}{3} < \frac{\varepsilon}{3} \,.$$

Außerdem gibt es, da wir eine CAUCHY-Folge haben, ein $N \geq M$, so daß $\left\|x_n - x_{n-1}\right\| < \frac{1}{3}\varepsilon$ für alle $n, m \geq N$. Damit ist

$$\begin{aligned}\|f(x^*) - x^*\| &= \left\|\big(f(x^*) - f(x_n)\big) + \big(f(x_n) - x_n\big) + \big(x_n - x^*\right\| \\ &\leq \|f(x^*) - x_{n+1}\| + \|x_{n+1} - x_n\| + \|x_n - x^*\| \\ &\leq \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon \,.\end{aligned}$$

Da dies für jedes $\varepsilon > 0$ gilt, muß $\|f(x^*) - x^*\| = 0$ sein, und das gilt nur, falls $f(x^*) = x^*$ ist. Damit haben wir einen Fixpunkt in X gefunden;

es gibt also genau diesen einen Fixpunkt, und für jeden Startwert x_0 konvergiert die Folge der $(x_n)_{n\in\mathbb{N}}$ gegen x^*.

■

f) Konvergenz von Funktionenfolgen

Im nächsten Kapitel werden wir zur Definition mehrdimensionaler Integrale auch Folgen von Funktionen betrachten. Zur Vorbereitung wollen wir hier bereits einige allgemeinere Tatsachen betrachten, für die kein Integralbegriff notwendig ist.

Seien zunächst $f_n\colon D \to \mathbb{R}$ irgendwelche Funktionen, die auf einer Teilmenge $D \subseteq \mathbb{R}^n$ definiert sind, und sei $f\colon D \to \mathbb{R}^m$ eine weitere Funktion. Für jedes $x \in D$ haben wir dann eine Folge $\big(f(x_n)\big)_{n\in\mathbb{N}}$ von Elementen aus $\mathbb{R}^n$ und können fragen, ob und gegebenenfalls wohin diese Folge konvergiert.

Definition: Eine Folge $(f_n)_{n\in\mathbb{N}}$ von Funktionen $f_n\colon D \to \mathbb{R}$ konvergiert auf der Teilmenge $A \subseteq D$ *punktweise* gegen die Funktion $f\colon D \to \mathbb{R}$, wenn für alle $x \in A$ gilt $\lim\limits_{n\to\infty} f_n(x) = f(x)$.

Als Beispiel betrachten wir die Folge der Funktionen $f_n\colon \mathbb{R} \to \mathbb{R}$ mit $f_n(x) = \cos^n x$. Für jedes $x \in \mathbb{R}$ haben wir dann die Folge $(\cos^n x)_{n\in\mathbb{N}}$. Falls $|\cos x| < 1$, ist das bekanntlich eine Nullfolge, falls $\cos x = 1$ haben wir die konstante Folge $(1^n)_{n\in\mathbb{N}}$, die gegen 1 konvergiert, und falls $\cos x = -1$, divergiert die Folge. Bekanntlich ist $\cos x = 1$ genau dann, wenn x ein ganzzahliges Vielfaches von 2π ist, und $\cos x = -1$ für alle ungeradzahligen Vielfachen von π. Setzen wir also

$$A = \mathbb{R} \setminus \big\{(2k+1)\pi \;\big|\; k \in \mathbb{Z}\big\},$$

so konvergiert die Folge der f_n auf A punktweise gegen die Funktion

$$f\colon \begin{cases} A \to \mathbb{R} \\ x \mapsto \begin{cases} 1 & \text{falls } x = 2k\pi \text{ für ein } k \in \mathbb{Z} \\ 0 & \text{sonst} \end{cases} \end{cases}.$$

Obwohl die Funktionen f_n allesamt stetig sind, ist die Grenzfunktion unstetig bei allen ganzzahligen Vielfachen von 2π; bei den ungeradzahligen Vielfachen von π existiert nicht einmal ein Grenzwert.

Wir können bei der Definition der Konvergenz aber auch anders vorgehen: Wir betrachten für eine beliebige Teilmenge $D \subset \mathbb{R}^n$ die Menge

$C_b(D,\mathbb{R})$ aller beschränkter Funktionen $f\colon D \to \mathbb{R}$, d.h. also die Menge aller Funktionen, für die es eine reelle Zahl M gibt, so daß $|f(x)| \leq M$ ist für alle $x \in D$.

Offensichtlich ist $C_b(D,\mathbb{R})$ ein Vektorraum, denn die Nullfunktion ist beschränkt, und für zwei beschränkte Funktionen f, g mit Schranken M, N und zwei reelle Zahlen a, b ist

$$|af(x) + bg(x)| \leq |a| \cdot |f(x)| + |b|\,|g(x)| \leq M\,|a| + N\,|b| \ .$$

Für jedes $f \in C_b(D,\mathbb{R})$ existiert

$$\|f\|_\infty \underset{\text{def}}{=} \sup\bigl\{\,|f(x)| \;\big|\; x \in D\bigr\}\,,$$

da die rechtsstehende Menge beschränkt ist. Wie bei der Maximumsnorm auf $\mathbb{R}^n$ rechnet man leicht nach, daß dies eine Norm auf $C_b(D,\mathbb{R})$ definiert, die sogenannte Supremumsnorm.

Wenn eine Folge $(f_n)_{n\in\mathbb{N}}$ von Funktionen aus $C_b(D,\mathbb{R})$ bezüglich dieser Norm gegen eine Funktion $f \in C_b(D,\mathbb{R})$ konvergiert, gibt es zu jedem $\varepsilon > 0$ ein $N \in \mathbb{N}$, so daß

$$\|f - f_n\|_\infty = \sup\bigl\{\,|f(x)| \;\big|\; x \in D\bigr\} < \varepsilon \quad \text{für alle } n \geq N\,;$$

insbesondere ist also für *jedes* $x \in D$ der Betrag von $f(x) - f_n(x)$ kleiner als ε für alle $n \geq N$. Das ist eine stärkere Eigenschaft als bei der punktweisen Konvergenz: Dort reicht es, wenn für jedes $\varepsilon > 0$ und jedes $x \in D$ ein $N \in \mathbb{N}$ existiert, so daß $|f(x) - f_n(x)| < \varepsilon$ für alle $n \geq N$; das N darf also von x abhängen.

Definition: Eine Folge $(f_n)_{n\in\mathbb{N}}$ von Funktionen $f_n\colon D \to \mathbb{R}$ konvergiert *gleichmäßig* gegen die Funktion $f\colon D \to \mathbb{R}$ auf der Menge $A \subseteq D$, wenn es zu jedem $\varepsilon > 0$ ein $N \in \mathbb{N}$ gibt, so daß $|f(x) - f_n(x)| < \varepsilon$ für alle $x \in A$ und alle $n \geq N$.

Die obige Folge der Funktionen $\cos^n x$ konvergiert nicht gleichmäßig gegen ihre Grenzfunktion: Andernfalls müßte sie insbesondere auf $\mathbb{R} \setminus \mathbb{Z}\pi$ gleichmäßig gegen die Nullfunktion konvergieren, d.h. zu jedem $\varepsilon > 0$ müßte es ein $N \in \mathbb{N}$ geben, so daß $|\cos^n x| < \varepsilon$ ist für alle $n \geq N$. Angenommen, es gäbe so ein N zu $\varepsilon = \frac{1}{2}$. Für $n \geq N$ wäre dann $|\cos^n x| < \frac{1}{2}$ für alle x, die keine ganzzahligen Vielfachen von π sind. Andererseits ist aber $|\cos^n x|$ eine auf ganz $\mathbb{R}$ stetige Funktion, die

bei allen ganzzahligen Vielfachen von π den Wert eins annimmt und somit in einer gewissen Umgebung dieser Punkte nur Werte annimmt, die größer sind als $\frac{1}{2}$.

Bei einer gleichmäßig konvergenten Folge kann es nicht passieren, daß die Grenzfunktion einer Folge stetiger Funktionen unstetig wird:

Lemma: $D \subseteq \mathbb{R}^n$ sei eine offene Menge und $f_n\colon D \to \mathbb{R}$ seien stetige Funktionen. Falls die Folge $(f_n)_{n\in\mathbb{N}}$ gleichmäßig gegen eine Funktion $f\colon D \to \mathbb{R}$ konvergiert, ist auch f stetig.

Beweis: Wir müssen zeigen, daß es zu jedem $x \in D$ und jedem $\varepsilon > 0$ ein $\delta > 0$ gibt, so daß $|f(y) - f(x)| < \varepsilon$ ist für alle $y \in D$ mit $\|y - x\| < \delta$, wobei $\|\cdot\|$ irgendeine Norm auf $\mathbb{R}^n$ bezeichnet.

Wegen der gleichmäßigen Konvergenz der Folge $(f_n)_{n\in\mathbb{N}}$ gibt es zunächst ein $N \in \mathbb{N}$, so daß $|f(y) - f_n(y)| < \frac{1}{3}\varepsilon$ ist für alle $n \geq N$ und alle $y \in D$. Wir wählen irgendein solches n. Wegen der Stetigkeit von f_n gibt es ein $\delta > 0$, so daß $|f_n(y) - f_n(x)| < \frac{1}{3}\varepsilon$ ist für alle $y \in D$ mit $|y - x| < \delta$. Für diese y ist dann auch

$$\begin{aligned} |f(y) - f(x)| &= \big|\big(f(y) - f_n(y)\big) + \big(f_n(y) - f_n(x)\big) + \big(f_n(x) - f(x)\big)\big| \\ &\leq |f(y) - f_n(y)| + |f_n(y) - f_n(x)| + |f_n(x) - f(x)| \\ &< \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon\,. \end{aligned}$$

Damit ist die Stetigkeit von f bewiesen. ∎